Functional Analysis of Collagen Galactosyltransferases and the Identification of Collagens in Giant Viruses

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Stephan Baumann

von Oberhallau SH

Promotionskomitee

Prof. Dr. Thierry Hennet (Vorsitz)

Prof. Dr. Lubor Borsig

Prof. Dr. Martin Hersberger

Prof. Dr. Richard Steet

Zürich 2016

CONTENT

1. **Introduction** 8
1.1 General introduction to collagens 8
1.2 Collagen biosynthesis 11
1.3 Posttranslational modifications 13
1.3.1 Prolyl Hydroxylation 13
1.3.2 Lysyl Hydroxylation 13
1.3.2 Glycosylation 14
1.3.3 Glycation 16
1.3.4 Collagen crosslinks 16
1.4 Biomedical application and production of recombinant collagen 17
1.4.1 Biomedical applications of collagen 17
1.4.2 Recombinant production of collagen 18
1.4.2.1 Collagen expression in Bacteria 18
1.4.2.2 Collagen expression in Yeast 19
1.4.2.3 Collagen expression in N. tabacum 20
1.4.2.4 Collagen expression in insect cells 20
1.4.2.5 Collagen expression in mammalian cells and transgenic animals 20
1.4.2.6 Comparison of recombinant human collagen expression systems 21
1.5 Bacterial and Viral collagens 22
1.5.1 Bacterial collagens 22
1.5.2 Viral collagens 23
1.6 Collagens in nucleocytoplasmic large DNA viruses (Mini-review) 24
1.6.1 Gene-structure analysis of viral collagen-like proteins 27
1.6.2 Non-collagenous domains 28
1.6.3 Collagenous domains 29
1.6.4 Evolutionary aspects of viral collagen-like proteins 31
1.6.5 Horizontal gene transfer 34
1.6.6 Conclusions 34
2. **Results** 38
2.1 Recombinant expression of hydroxylated human collagen in *Escherichia coli* 39
2.1.1 Abstract 40
2.1.2 Introduction 41
2.1.3 Materials and Methods 43
2.1.4 Results 48
2.1.5 Discussion 53

2.2 Collagen accumulation in osteosarcoma cells lacking GLT25D1 collagen galactosyltransferase 70
2.2.1 Abstract 71
2.2.2 Introduction 72
2.2.3 Materials and Methods 74
2.2.4 Results 78
2.2.5 Discussion 83
3. **Discussion** 102
3.1 Conclusions 102
3.2 Future directions 107
4. **References** 109
5. **Acknowledgements** 119
6. **Curriculum Vitae** 120

Summary

Collagens are structural proteins consisting of $(Gly\text{-}X\text{-}Y)_n$ repeat units and have been identified in vertebrates, invertebrates, bacteria, fungi and viruses. In mammals, collagens occur in connective tissues, bones and tendons, in which they confer stability and tensile strength. In tendons, collagen fibrils associate to collagen fibers, which are the structural components of collagen fascicles. Collagen fibrils are made of three collagen α-chains, a structural entity that results in the tremendous strength of collagen. Even though collagens were identified more than a century ago, several molecular properties have remained unexplored. Human collagens are posttranslationally modified by prolyl- and lysyl hydroxylation and the subsequent glycosylation of hydroxylysine residues. While prolyl hydroxylation provides stabilization of the triple helical structure of collagen, lysyl hydroxylation can be additionally modified by the lysyl-oxidase enzyme family to crosslink collagen molecules. The function of collagen glycosylation however is unknown.

β(1-O)galactosylation of collagen is conferred by the two paralogous enzymes GLT25D1 and GLT25D2 in the endoplasmic reticulum. While *GLT25D1* is expressed ubiquitously, *GLT25D2* is restricted to the brain and skeletal muscles. The galactosyl residue can be further elongated with glucose to the Glc(α1-2)Gal(β1-O)Hyl disaccharide. The responsible glucosyltransferase has not been identified yet. Several functions have been assigned to collagen glycosylation including folding, secretion and receptor interactions with integrins or the urokinase-type plasminogen activator receptor associated protein. No study however has addressed the function of GLT25D1 or GLT25D2.

In this thesis, we used the CRISPR/Cas9 system to inactivate *GLT25D1* and *GLT25D2* in osteosarcoma cell lines. Glycosyltransferase activity was significantly reduced in *GLT25D1*-null cells to 3 – 7% of the activity measured in wild type cells. Interestingly, collagen glycosylation in *GLT25D1*-null cells as was measured by HPLC amino acid analysis was only reduced by 40 – 60% indicating a potential compensation mechanism. We found a threefold upregulation of *GLT25D2* in *GLT25D1*-null cells by quantitative PCR analysis. We further found increased collagen type I expression levels in *GLT25D1*-null cells but not in *GLT25D2*-null cells as was assessed by quantitative PCR analysis, Western blot analysis and immunofluorescence. Expression levels of the collagen types III and V remained unchanged. In contrast to previous studies, we detected normal collagen secretion and folding as measured by a pulse chase experiment and circular dichroism.

In the second part of the thesis, we identified and analyzed collagens from nucleocytoplasmic large DNA viruses *in silico*. Collagen-like genes have previously been described in the *Acanthamoeba polyphaga* mimivirus. Based on the recent discovery of further members of the *Mimiviridae* family, we identified 142 collagen-like genes in the genome of 60 nucleocytoplasmic large DNA viruses. Analysis of the

collagen-like proteins revealed distinct phylogenetic origins for *Pithoviridae* and *Pandoraviridae* but a shared phylogenetic origin for *Mimiviridae* and *Megaviridae*. Sequence analysis of the viral collagen-like proteins hint at three distinct mechanisms for triple helix stabilization. Most probably, *Pandoraviridae* use prolyl hydroxylation at the Y position of Gly-X-Y repeats to stabilize collagen folding, similar to animal collagens. *Pithoviridae* in contrast incorporate threonine at the Y position of 50% of all Gly-X-Y repeats, indicating a similar stabilization mechanism as is found in the deep-sea worm *Riftia pachyptila*, which uses glycosylated threonine residues to increase triple-helical stability. *Mimiviridae* and *Megaviridae* show a unique pattern of oppositely charged amino acids at the X and Y position, which could result in electrostatic stabilization of the collagen triple helix. These models were generated *in silico*, however, experimental work is still required in order to characterize the collagen-like proteins from giant viruses and to support the interpretation of our data.

In the third part of the thesis, we used viral enzymes to produce posttranslationally modified collagens in *E. coli*. Giant viruses not only encode collagen-like proteins but also collagen modifying enzymes. Previously, a viral prolyl 4-hydroxylase and a bifunctional collagen lysyl hydroxylase and glucosyltransferase were identified in the genome of the mimivirus. We expressed human collagen together with these viral enzymes and showed posttranslational modifications at lysine and proline residues resulting in improved triple helical stability compared to unmodified collagen. We further used the recombinant collagen as an extracellular matrix substitute for the growth of human umbilical vein cells proving biocompatibility of the recombinant human collagen. In the future, this expression system could be used for the large-scale production of recombinant human collagen suitable for a variety of biomedical applications.

In conclusion, we contributed to the general understanding of the diversity of collagens and collagen modifications. The results will be beneficial for further studies on collagen glycosylation in more complex model systems. The *in silico* studies on viral collagens help resolving the phylogeny of collagens and sets a fundamental basis for experimental work to identify the functional role of viral collagens.

Zusammenfassung

Kollagene sind strukturelle Proteine, die durch den Baustein der Aminosäuren $(Gly\text{-}X\text{-}Y)_n$ gekennzeichnet sind und in Vertebraten, Invertebraten, Bakterien, Pilzen und in Viren vorkommen. Bei Säugetieren kommt Kollagen in Bindegewebe, Knochen und Sehnen vor, in welchen sie durch ihre verdrillte Struktur für die Zugfestigkeit und die Stabilität des Gewebes verantwortlich sind. Sehnen sind zum Beispiel aus Faszikeln aufgebaut, welche aus Kollagenfasern bestehen, die wiederum in Kollagenfibrillen unterteilt sind. Letztere sind durch drei Kollagen α-Ketten charakterisiert, ein strukturelles Konstrukt, das zur enormen Stabilität des Gewebes führt. Obwohl Kollagene schon vor über einem Jahrhundert beschrieben worden sind, sind einige molekulare Eigenschaften noch immer unerforscht. Humane Kollagene sind posttranslational modifiziert mit Prolyl- und Lysyl Hydroxylierung und anschliessender Glykosylierung von spezifischen Hydroxylysin Seitenketten. Prolyl Hydroxylierung steigert die Stabilität der Kollagen Tripelhelix. Lysyl Hydroxylierung dient den Lysyl Oxidase Enzymen als Substrat, welche die verschiedenen Kollagenketten kovalent miteinander verknüpfen. Die physiologische Funktion der Kollagenglykosylierung ist allerdings unbekannt.

Der Kern der Kollagenglykosylierung besteht aus β(1-O)Galaktose und wird von den Enzymen GLT25D1 und GLT25D2 im endoplasmatischen Retikulum übertragen. *GLT25D1* wird in den meisten Zellen ubiquitär exprimiert. Im Gegensatz dazu wird *GLT25D2* lediglich im Gehirn und in kleineren Mengen in Skelettmuskelzellen exprimiert. Die Galaktose auf Hydroxylysin kann mit Glukose zum Disaccharid Glc(α1-2)Gal(β1-O)Hyl erweitert werden. Obwohl die Hauptfunktion der Kollagenglykosylierung unbekannt ist, wurde die Beteiligung dieser posttranslationalen Modifikation in einigen Studien geprüft. Demnach spielt die Glykosylierung in der Faltung und in der Sekretion von Kollagenen sowie in der Interaktion von Kollagenen mit Rezeptoren wie zum Beispiel Integrinen oder dem Urokinase-Typ Plasminogen assoziierten Rezeptor Protein eine Rolle. Keine Studie erforschte bisher die Funktionen von GLT25D1 oder GLT25D2.

In dieser Dissertation verwendeten wir das CRISPR/Cas9 System um *GLT25D1* und *GLT25D2* in Osteosarkomazellen zu inaktivieren. Die Glykosyltransferaseaktivität war in *GLT25D1*-null Zellen signifikant um 93 – 97% reduziert. Jedoch detektierten wir bei Messungen von der Aminosäurezusammensetzung in Kollagenen von *GLT25D1*-null Zellen weiterhin 40 – 60% der ursprünglichen Kollagenglykosylierung. Wir fanden eine dreifache Überregulierung von GLT25D2 in den GLT25D1-null Zellen, gemessen mit einer quantitativen PCR Analyse, welche die restliche Glykosylierung erklären könnte. Des Weiteren entdeckten wir erhöhte Kollagen Typ I Expressionslevel gemessen mit quantitativer PCR Analyse, mit Western Blot Analyse und mit Immunfluoreszenz. Die Expressionslevel von

Kollagen Typ III und V blieben unverändert. Anders als in vorhergehenden Studien berichtet wurde, konnten wir keinen Unterschied bei der Kollagensekretion oder der Faltung feststellen.

Im zweiten Teil der Dissertation identifizierten und analysierten wir Kollagene von grossen nucleocytplasmatischen DNS Viren *in silico*. Kollagene wurden bereits im Genom vom *Acanthamoeba polyphaga* Mimivirus nachgewiesen. Aufgrund der Entdeckung von weiteren Mitgliedern der Familie der *Mimiviridae*, suchten wir in 60 Mitgliedern der grossen nucleocytoplasmatischen DNS Viren nach Kollagensequenzen und entdeckten 142 kollagenähnliche Proteine hauptsächlich in den Familien der *Mimiviridae*, *Megaviridae*, *Pandoraviridae* und im Pithovirus. Die phylogenetische Analyse von den kollagenähnlichen Proteinen zeigte drei verschiedene Abstammungen der Kollagene vom Pithovirus, den *Pandoraviridae* und einen gemeinsamen Ursprung der *Mimiviridae* und *Megaviridae*. Die Analyse der Aminosäurezusammensetzung der Kollagendomänen der kollagenähnlichen Proteine deutete auf drei unterschiedliche Mechanismen zur Tripelhelix Stabilisierung hin. Dabei inkorporieren Pithoviren in 50% der Y Position von den Gly-X-Y Einheiten Threonin, ähnlich wie der Tiefseewurm *Riftia pachyptila*. Dieser Tiefseewurm benutzt glykosylierte Threonine zur Stabilisierung der Tripelhelix. Pandoraviren besetzen die X und Y Position vor allem mit Prolin, ähnlich wie menschliche Kollagene. Im Gegensatz hierzu beinhalten kollagenähnliche Proteine der *Mimiviridae* und *Megaviridae* vor allem geladene Aminosäuren in der X und Y Position der Gly-X-Y Einheiten. Dies weist auf eine elektrostatische Stabilisierung der Kollagen Tripelhelix hin. Weil diese Ergebnisse *in silico* generiert wurden, werden experimentelle Arbeiten benötigt, um die kollagenähnlichen Proteine detailliert zu charakterisieren und um die Interpretation unsere Daten zu unterstützen.

Im dritten Teil der Dissertation verwendeten wir virale Enzyme um posttranslational modifiziertes Kollagen in *E. coli* zu exprimieren. Grosse Viren besitzen nicht nur Kollagene, sondern auch Enzyme zur Modifikation der Kollagene. Frühere Studien identifizierten eine Prolyl 4-Hydroxylase und eine bifunktionelle Kollagen Lysyl Hydroxylase und Glukosyltransferase im Mimivirusgenom. Wir exprimierten humanes Kollagen Typ III zusammen mit diesen Enzymen und wiesen die posttranslationalen Modifikationen im Kollagen mittels HPLC-gestützter Aminosäurenanalyse nach. Die Modifikationen resultierten in einer höheren Stabilität verglichen mit unmodifiziertem Kollagen. Des Weiteren benutzten wir das rekombinante Kollagen zur Kultivierung von menschlichen Nabelschnurzellen und bewiesen dadurch dessen Biokompatibilität. Dieses Expressionssystem könnte in Zukunft für die Massenproduktion von rekombinanten Kollagenen für eine Vielfalt von biomedizinischen Anwendungen gebraucht werden.

Zusammenfassend trugen wir mit dieser Dissertation zum Verständnis der Vielfalt von Kollagenen und kollagenmodifizierenden Enzymen bei. Die Resultate werden zukünftige Studien über Kollagenglykosylierung in komplexeren Organismen wie in Mäusen oder Zebrafischen unterstützen. Die Studie über virale Kollagene hilft zudem, die Herkunft und die funktionelle Rolle besser zu verstehen.

List of Abbreviations

ER	Endoplasmic reticulum
FCS	Fetal calf serum
Gal	Galactose
Glc	Glucose
Gly	Glycine
gRNA	guide RNA
Hyl	Hydroxylysine
LH3	Lysyl hydroxylase 3 (protein)
NCLDV	Nucleocytoplasmic large DNA viruses
PLOD3	Lysyl hydroxylase 3 (gene)

1.INTRODUCTION

1.1 GENERAL INTRODUCTION TO COLLAGENS

Collagens are a superfamily of extracellular matrix proteins serving as connective elements in various tissues in vertebrates and invertebrates. Collagens present the most abundant protein in the animal kingdom with 25 – 30% of all body proteins [2]. Collagens share a common repeated structural motif of $(Gly\text{-}X\text{-}Y)_n$ which is however not confined to collagens but also occur in other proteins, such as the mannose binding lectin or adiponectin [3]. In humans, 28 different collagen types have been identified which are encoded by more than 40 genes. Based on their quaternary structure and their function, they can be assigned to families of fibrillar collagens, fibril-associated collagens, network forming basement membrane collagens and transmembrane collagens (Table I) [4], [5]. Generally, collagens are modified by prolyl 3- and prolyl 4-hydroxylation, lysyl hydroxylation, and hydroxylysine O-glycosylation. These posttranslational modifications are essential for collagen folding, secretion, and for structural properties such as the tensile strength and tenacity as discussed later.

Table I: human collagens and collagen-like proteins, adapted from [6].

Sub-family	Members
Fibrillar collagens	Types I, II, III, V, VI, XI, XXIV, XXVII
Fibril associated collagens	Types IX, XII, XIV, XVI, XIX, XX, XXI
Basement membrane collagens	Type IV, VII, XV, XVIII
Transmembrane collagens and collagen-like proteins	Types XIII, XVII, XXIII, XXV, ectodysplasin, macrophage scavenger receptors I – III, MARCO, SRCL, gliomedin, CL-P1
Collectins and ficolins	Mannan binding lectin, surfactant protein A and D, conglutin, CL-43, CL-16, CL-L1, CL-P1, L-, M- and H-ficolins
Other collagen-like proteins	Emu1 and 2, acetylcholinesterase tail subunit, adiponectin

The family of fibrillar collagens consists of collagens type I, II, III, V, XI, XXIV and XXVII [7-9]. In humans, the most prevalent collagen is the fibrillar collagen type I, which presents up to 90% of the total collagen content. Collagen type I is found in tendons, skin, artery walls, cornea, fibrocartilage and in the organic part of bones and teeth. Up to 40% of collagen type I is found in the skin [10]. The second most prevalent collagen is collagen type II, which is restricted to cartilage tissue. Collagen type III is part of the granulation tissue and its expression is induced after tissue injury. It also is associated with collagen type I in skin and

supports collagen I fibrillogenesis in cardiovascular development [11]. The other fibrillar collagens are only found in minor amounts and their functions are poorly investigated. Some of these collagens are associated with either collagen type I or type II where they modify fibril diameter (collagen type XI) or help initiate collagen fibrils to assemble (collagen type V) [12, 13].

Collagens have a unique structure, which is described in Figure 1. Mature collagen, as found in tendons, occurs in form of collagen fascicles, which are a composition of many collagen fibrils. Fibrils are composed of the procollagen triple helix, which consists of three procollagen molecules. The prototypical fibrillar collagen molecule carries a non-collagenous *N*-terminal propeptide domain, an *N*-terminal telopeptide, a collagenous domain, a *C*-terminal telopeptide and a non-collagenous *C*-terminal propeptide. The *N*- and *C*- terminal propeptides are cleaved before supramolecular assembly (See chapter 1.2) and collagen molecules are crosslinked by the enzymes from the family of lysyl oxidases to stabilize the collagen fibrils.

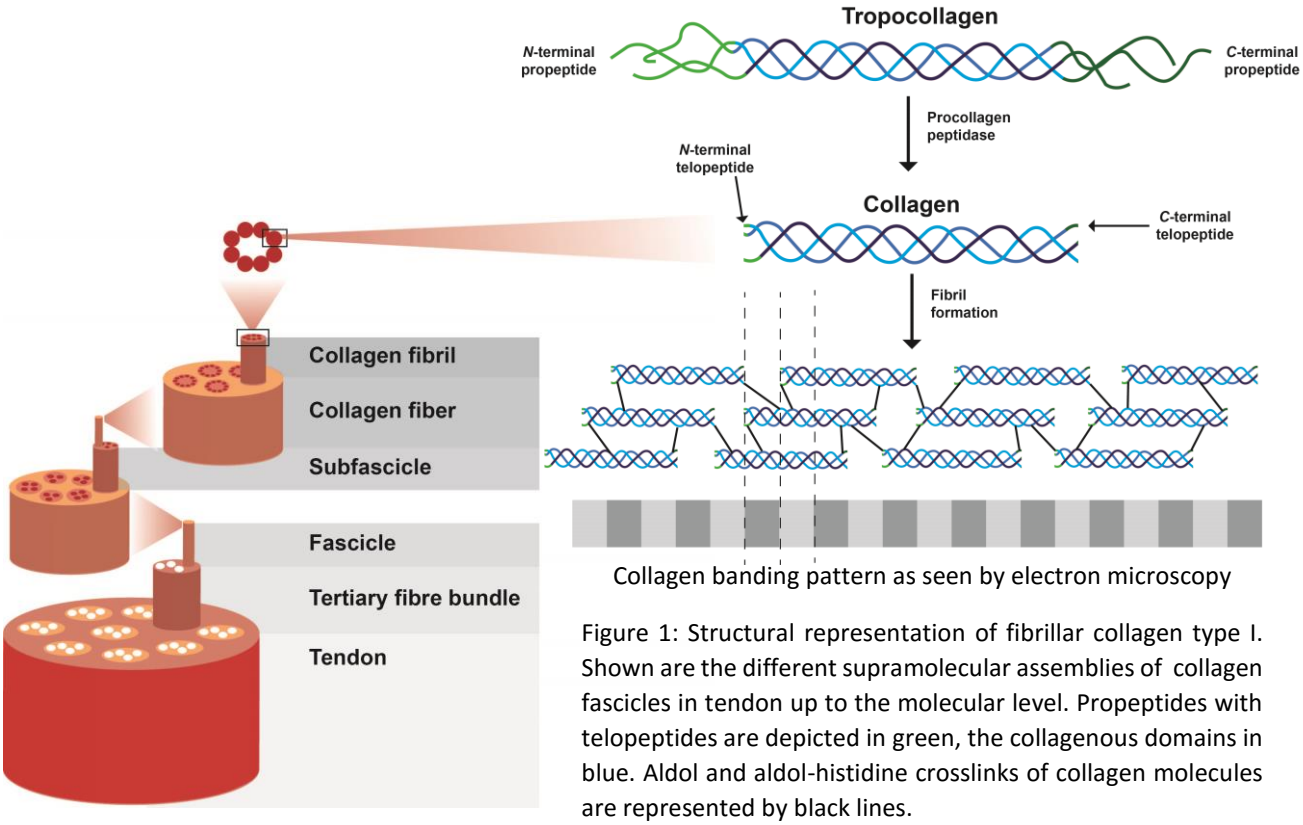


Figure 1: Structural representation of fibrillar collagen type I. Shown are the different supramolecular assemblies of collagen fascicles in tendon up to the molecular level. Propeptides with telopeptides are depicted in green, the collagenous domains in blue. Aldol and aldol-histidine crosslinks of collagen molecules are represented by black lines.

The major part of fibrillar collagens is represented by the collagenous domains whose general amino acid composition consists of Gly-X-Y repeats ranging from 154 repeats in collagen type VIII up to 490 repeats in collagen type VII. The small amino acid glycine is a prerequisite for the tight packaging of three collagen molecules in a polyproline type II triple helix. Glycine enables a helical pitch of 10/3 to 7/2. This small pitch in turn allows interchain hydrogen bonds between glycine and proline to stabilize the triple helix and regulate the triple helix density [14]. The three-dimensional structure of the collagen helix is a prerequisite for interaction of collagens with other biomolecules such as integrins or heparin [15-17].

Numerous diseases, for instance Ehlers-Danlos syndrome and osteogenesis imperfecta, are associated with interruptions of the Gly-X-Y pattern caused by mutation of glycine into any other amino acid [18, 19]. Collagenous domains are further characterized by a high proline content at position X and Y of the Gly-X-Y repeats. In human collagens approximately 22% of all residues are proline or hydroxyproline [20]. Due to the pyrrolidin ring of proline, procollagen strands pre-align in a helix-like manner and the entropic cost for triple helix formation is thereby reduced [21]. Many of the proline residues in the Y position are hydroxylated. Hydroxylation is conferred by the prolyl hydroxylases and stabilizes the triple helix essentially [22, 23] (see chapter 1.3.1).

The collagenous domain of fibrillar collagens is interjacent of two non-collagenous propeptides. These propeptides prevent premature fibrillogenesis and are in most cases absent in mature collagen fibrils. Once the procollagen is secreted from the cell, the propeptides are removed by specific metalloproteases. Nevertheless, they play a crucial role in intracellular collagen folding, trimerization and secretion (see chapter 1.2). A collagen triple helix can consist either of three identical chains as homo-trimer or of three different chains as hetero-trimer. Since several procollagen molecules can be part of different types of collagen, the *C*-terminal propeptides serve as nucleation site and are the main factor of chain selectivity. Even though the human *C*-terminal propeptides of fibrillary collagens share a relatively high sequence homology of 46% [24], they build a distinct electrochemical environment allowing only specific procollagens to assemble [25]. Once the initial contact of three C-terminal propeptides is established, the collagen triple helix forms auto catalytically along the collagenous domain towards the *N*-terminal propeptide, driven by enthalpy *via* hydrophobic interactions [26]. The *C*- as well as the *N*-terminal propeptide contain cysteine residues that are intramolecularly crosslinked by disulfide bridges. Contrary to the propeptides, the small telopeptides remain attached to the collagenous domains after fibril formation (Fig. 1). They contain several hydroxylysine residues, which are intermolecularly cross-linked via lysyl oxidase enzymes [27].

Most of the non-fibrillar collagens belong to the family of fibril-associated collagens with interrupted triple helices (FACIT). In comparison to fibrillar collagens, they contain several triple helical domains interrupted by non-collagenous domains and do not assemble into fibers (Fig. 2). FACIT collagens are associated with collagen fibrils and are low abundant (5 - 15% of cartilage collagens). The FACIT family comprises collagens type IX, XII, XIV, XVI, XIX, XX, XXI, and XXII. Collagen type XII and XIV are associated with type I collagen, collagen type IX and XVI with type II collagen (Fig. 2). These collagens have various functions, which are poorly investigated [28]. Since the propeptides are not cleaved in non-fibrillar collagens, they can act as anchor or linker of various fibrillar collagens [29]. Type IX collagen is further

attached to a glycosaminoglycan chain strengthening the collagen fibrils additionally. Mutations in the collagen type IX lead to osteoarthritis [30] indicating non-functional cartilage composition.

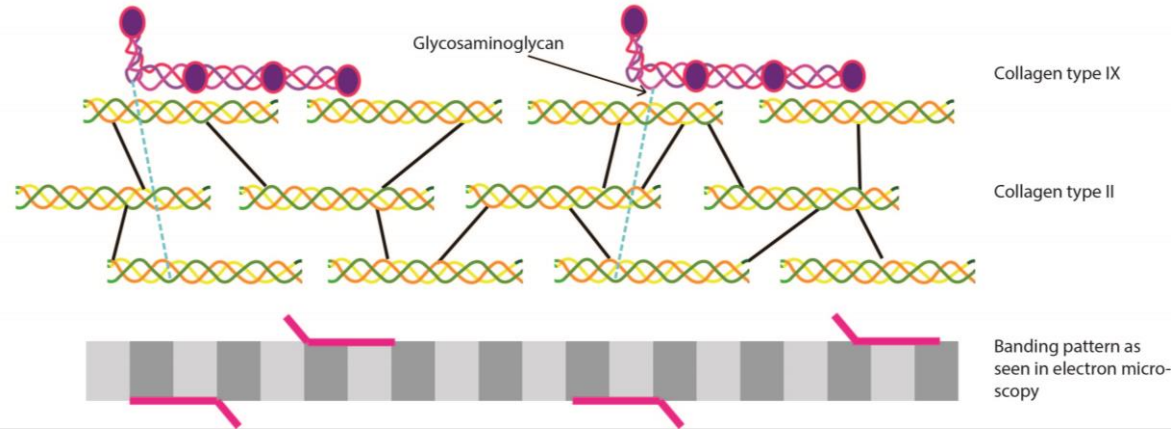


Figure 2: Structural representation of FACIT collagens. Collagen type IX (purple) with non-collagenous domains (violet ovals) associated to collagen type II (yellow). Collagens are linked to each other via aldol and aldol-histidine linkages (black). Collagen type IX is attached to a glycosaminoglycan chain (cyan).

The most prominent member of the basement membrane type collagens is the non-fibrillar collagen type IV. It builds a network like lattice and is part of the basement membranes. Collagen type IV is associated with a variety of non-collagenous extracellular matrix proteins such as laminins, nidogens and the proteoglycan perlecan. Other non-fibrillar collagens such as type VII and XVII are also part of basal membranes and anchor epithelial cells. Mutations in type IV collagen result in the Alport syndrome [31], mutations in type VII collagen in epidermolysis bullosa [32], a disease originating from missing anchorage of the epithelium to the basement membrane.

1.2 COLLAGEN BIOSYNTHESIS

Collagen biosynthesis is a complex process composed of intracellular and extracellular steps (Fig. 3). The mechanisms are not yet fully understood and the steps that have been identified are mostly related to fibrillary collagens only. Pre-procollagen chains are translated in the rough endoplasmic reticulum (ER) where they undergo several posttranslational modifications such as prolyl and lysyl hydroxylation and N- and O-linked glycosylation [33]. Still in the ER, the *C*-terminal signal peptide is cleaved off and three monomeric pro-collagens associate at the *C*-terminal propeptides. The *C*-terminal propeptides are responsible for the monomer association and chain selectivity. Four cysteine residues in each *C*-terminal propeptide are intramolecularly linked via disulfide bonds formed by the protein disulfide isomerase. The associated propeptides are further intermolecularly linked by another six disulfide bonds. Propeptide association initiates a zipper like trimerization of the three procollagen molecules from the *C*- to the *N*-terminus. The trimerization process is controlled by a peptidyl-prolyl-cis-trans-isomerase and is assisted by the molecular chaperones HSP47, GRP78 and Grp94. These chaperones bind incorrectly folded procollagen and prevent premature secretion [34]. Furthermore, the *N*-terminal signal peptide is cleaved

off before the ER resident integral membrane protein Tango1 targets the procollagen for COPII coated transport to the Golgi apparatus [35]. However, Tango1 only targets collagen type VII and is not involved in collagen type I secretion [35]. The mechanism of targeting collagen type I for the transport to the Golgi apparatus is unknown.

Regular COPII coated transport vesicles are restricted to molecules < 100 nm and hence not suitable for transporting 300 nm procollagens. The COPII coat consists of a structured grid composed of the proteins SEC13 and SEC31. A small pool of the SEC31 proteins can be ubiquitinylated by the ubiquitin ligase CUL3-KLHL12 [36]. This posttranslational modification results in a structural change in the COPII coat and induces the formation of vesicles with an increased diameter of up to 500 nm, which is large enough to accommodate procollagen molecules [37]. The exit route of collagen from the trans Golgi network to the cell surface is unknown [38].

After or during the late phase of the transport to the outside of the cell, the procollagen molecules are further modified. This step involves trimming of the *N*-terminal propeptide by ADAMTS2 [39], as well as trimming of the *C*-terminal propeptide by the procollagen *C*-terminal propeptide proteinase or tolloid-like proteinases [40], depending on the type of collagen. Removal of the propeptides is essential for self-assembly of the collagen molecules into larger collagen fibrils [41]. The final step in collagen biosynthesis includes extracellular covalent cross-linking of collagen molecules by the family of lysyl oxidases. These proteins catalyze the condensation of two lysine or hydroxylysine residues, located at the telopeptide regions of the collagens [42].

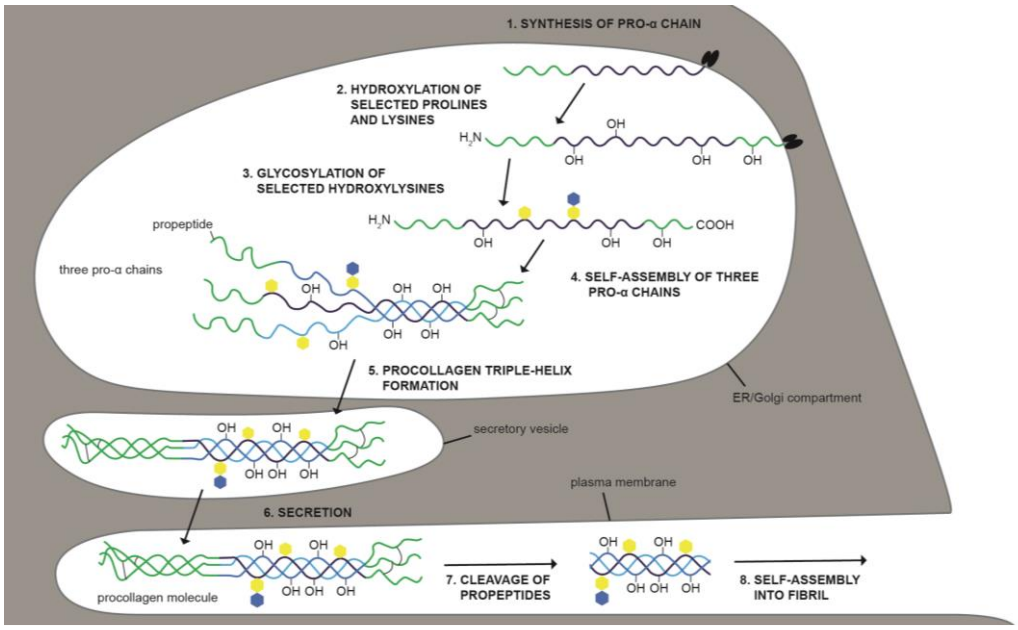


Figure 3: Collagen biosynthesis. 1. Procollagens are translated in the rough endoplasmic reticulum. 2. They are cotranslationally hydroxylated and 3. selected lysine residues are glycosylated. 4. Procollagens assemble at the *C*-terminus and are covalently connected *via* disulfide bridges. 5. Procollagen formation is completed and the molecule is transported outside the cell. 7. Propeptides are cleaved off. 8. Procollagens assemble into fibrils before they are covalently linked via the lysyl oxidase enzyme. Figure is adapted from [1]

1.3 POSTTRANSLATIONAL MODIFICATIONS

1.3.1 PROLYL HYDROXYLATION

The most abundant posttranslational modification of collagen is prolyl 4-hydroxylation presenting up to 38% of the total amino acid composition of collagens [20] and around 4% of all amino acids found in animal tissue. The reaction is catalyzed by the tetrameric prolyl 4-hydroxylase with prolyl-procollagen, 2-oxoglutarate and elemental oxygen as substrates. Fe(II) and ascorbate are essential cofactors for prolyl 4-hydroxylation. The prolyl 4-hydroxylase is mainly expressed in the lumen of the endoplasmic reticulum and is essential for the triple-helical stability of collagen at body temperature of warmbloods. The effect of this posttranslational modification becomes evident by comparison of the melting temperature of fully hydroxylated collagen type I (T_m = 43°C) with the unhydroxylated form (T_m = 27°C) [22]. Other than previously thought, the stabilizing effect does not origin from hydrogen bonding of the hydroxyl group of collagen strands but from stereoelectronic effects mediated by the hydroxyl oxygen [43].

Prolyl 4-hydroxylase modifies procollagen and around 15 other non-collagenous proteins. While a polyproline type II fold is recognized by the prolyl 4-hydroxylase, polyproline alone is not hydroxylated by the enzyme [44]. The minimum substrate required for hydroxylation is Gly-X-Pro with proline preferred also at position X. Substrate affinity increases with length of the substrate [45]. Mutations in the prolyl 4-hydroxylase β-subunit protein disulfide isomerase results in the Cole-Carpenter syndrome and is manifested by fragile bones and skull deformations [46].

Another form of hydroxylation is conferred by the prolyl 3-hydroxylase. 3-hydroxyproline is much less abundant than 4-hydroxyproline and is more frequent in collagen types IV and V than in collagen type I or II with 10 - 15 compared to a single 3-hydroxyproline residue per molecule [47]. 3-hydroxyproline formation occurs in the X position of Gly-X-4-Hyp [48] and is conferred by one of three isoforms of prolyl 3-hydroxylases. Different to prolyl 4-hydroxylation, prolyl 3-hydroxylation only affects triple-helical stability marginally [49] and must hence possess other functions that remain unknown up to date. Mutation in the prolyl 3-hydroxylase gene *LEPRE1* results in a disorder resembling lethal osteogenesis imperfecta [50].

1.3.2 LYSYL HYDROXYLATION

Hydroxylysine accounts for about 5‰ of all amino acid residues in whole skin collagen [51]. Equal to the previously described prolyl hydroxylases, lysyl hydroxylation requires Fe(II), oxygen, 2-oxoglutarate and ascorbate for enzymatic activity. In contrast to the soluble tetrameric prolyl 4-hydroxylases, the lysyl

hydroxylases are ER membrane bound and homodimeric. The lysyl hydroxylases only act on native collagen chains before triple helix formation. In humans, there exist three isoforms of lysyl hydroxylases encoded by the genes *PLOD1*, *PLOD2* and *PLOD3*. All three isoforms are expressed in the same cells in skin, lung, placenta, cartilage, spleen, brain and liver [52] but could not be correlated to expression of a specific type of collagen [53] indicating unspecific activity towards various types of collagen. Even though many types of collagen can be hydroxylated by all three isoforms, the lysyl hydroxylases contain specificity towards the location of the lysines in collagen. PLOD1 hydroxylates preferably lysines located in the collagenous domains, while PLOD2 hydroxylates lysine residues in the telopeptide region [54, 55]. Such a specificity is not known for PLOD3.

Lysyl hydroxylation is a prerequisite for two further posttranslational modifications of collagen. Lysine residues located in the telopeptide region of collagen are hydroxylated before being covalently crosslinked to form intra- and intermolecular bonds by the lysyl oxidase enzymes. Lysine residues mainly located in the triple helical region of the collagens are converted to hydroxylysine to serve as acceptor for collagen galactosyltransferases. Mutations of any of the three isoforms of lysyl hydroxylases result in severe connective tissue disorders. Mutations in PLOD1 result in Ehlers-Danlos syndrome VI [56], mutations in PLOD2 in the Bruck syndrome [57] and mutations in PLOD3 in a connective tissue disorder combining symptoms of osteogenesis imperfecta and the Ehlers-Danlos syndrome [58].

1.3.2 Glycosylation

Two distinct types of O-glycosylation have been identified in collagen. Hydroxylysine residues in collagen can either be monoglycosylated with galactose or diglycosylated with the disaccharide glucosylgalactose (Glc(α1-2)Gal(β1-O)Hyl) [59]. Core glycosylation by galactose is transferred by either GLT25D1 or GLT25D2 [60]. While *GLT25D1* is expressed ubiquitously in all tissues, *GLT25D2* expression is limited to brain and skeletal muscles. GLT25D1 and GLT25D2 are soluble ER resident enzymes [61] transferring galactose from UDP-galactose to collagenous hydroxylysine residues but not to free hydroxylysine [62]. The main glucosyltransferase transferring glucose from UDP-glucose to galactosyl-hydroxylysyl residues is currently unknown. The multifunctional lysyl hydroxylase enzyme LH3 was shown to transfer glucose to galactosylated collagen [63] in low amounts. Glucosylated collagen however also occurs in sponges or chicken, which lack a LH3 homologue. It is hence likely that the main collagen glucosyltransferase in humans has not yet been identified.

Collagen glycosylation takes place on nascent collagen strands but not on triple helical collagen [64, 65]. The extent of collagen glycosylation differs in tissues and types of collagen. Collagen type IV is believed

to be the most glycosylated collagen, whereas fibrillary collagens type I and II carry less collagen glycosylation [66].

Many biological roles have been assigned to collagen glycosylation. Most studies were performed on PLOD3 knockout mice and hence leave collagen galactosylation unchanged. Mutation of the glucosyltransferase domain of LH3 results in embryonal lethality at E9.5 in transgenic PLOD3 mice possibly by altered secretion and assembly of collagen type IV. Fibroblasts extracted from the embryos showed aggregated collagen type IV at the ER, which resulted in dilated ER and a condensed Golgi. There occured no tetramerization of collagen type IV in KO fibroblasts and secretion was heavily impaired [67]. A similar observation was found in *C. elegans* let-268 knockout worms [68]. Let-268 is a *PLOD3* homolog (56% sequence identity) and hydroxylates lysine residues. Glycosylation activity however has not been identified in let-268 yet. Let-268 knockout worms show deficient collagen type IV secretion and lack a basal membrane. All these experiment involving the bifunctional lysyl hydroxylase 3 base on alteration of the glucosyltransferase activity. However, mutations in lysyl hydroxylase 3 glycosyltransferase domain probably affect its lysyl hydroxylase activity as well. It is hence difficult to assess the role of the collagen glucosyltransferase alone. Nevertheless, these results suggest that collagen glycosylation plays a crucial role in collagen folding and secretion and further influences supramolecular assembly of collagen IV.

Other experiments used chemically deglycosylated collagens and mimic the physiological situation only up to a limited extent. The urokinase-type plasminogen activator receptor associated protein uPARAP internalizes collagen dependent on its fibronectin II domain [69]. The uptake efficiency has been shown to be influenced by its lectin domain [70]. uPARAP however does not only mediate uptake of the highly glycosylated collagens type IV and V but also of the less glycosylated collagen type I, indicating a minor influence of collagen glycosylation [71].

Several other studies described a steric role for collagen glycosylation. Due to the bulky and polar nature of glycans, they can modulate collagen receptor interactions. It was shown that glycosylation disturbs melanoma α2β1 and α3β1 integrin interactions with type IV collagen [72]. Another study suggested a correlation of collagen type I glycosylation and the valence of crosslinks in fibrillary collagens [73] possibly by steric hindrance of the lysyl oxidase enzymes. Hyperglycosylated collagen fibrils from an osteogenesis imperfecta mouse model showed an increased fibril diameter [74].

Even though collagen glycosylation has been studied for decades in various biological ways, no conserved mechanism or function was identified. Therefore, further studies involving the collagen galactosyltransferases GLT25D1 and GLT25D2 are inevitable.

1.3.3 Glycation

Carbohydrates cannot only be covalently bound to collagen via glycosyltransferases, but can also be spontaneously attached via non-enzymatic glycosylation (glycation). The sugars glucose, fructose or galactose are typically involved in glycation. Glycation efficiency however is ten times higher when galactose or fructose is attached in comparison to glucose [75]. Due to the usually low availability and the slow reactivity of glucose in blood, glycation in a pathological extent only occurs in diabetic individuals with prolonged hyperglycemia or in aged individuals [76]. Non-enzymatic glucose addition to proteins occurs spontaneous via formation of a labile Schiff base on available amino groups of lysine residues. Over time, the Schiff base rearranges into a more stable ketoamine product and is thus irreversibly linked to the protein. Glycated proteins can further undergo various modifications resulting in advanced glycation endproducts [77]. Due to the long half-life of collagens, they are the most likely site for glycation to occur. After years of elevated blood sugar levels or in aged healthy individuals, these modifications accumulate and lead to functional deficits of connective tissue and wound healing [78, 79]. The stiffening of the tissue can be explained by crosslinking of advanced glycosylation endproducts of various proteins such as collagen or elastin and modification of the molecular properties of membranes and cell surfaces. On the surface of many cell types, advanced glycosylation endproducts can be sensed via receptors [80]. The receptors for advanced glycosylation endproducts then activate pro-inflammatory pathways via NF-κB. NF- κB upregulates the receptors and establishes an inflammatory positive feedback cycle that can result in organ damage or even organ failure [81]. Many diseases such as atherosclerosis, myocardial infarction, diabetic retinopathy, neuropathy and nephropathy are linked to glycated collagens [82-84].

1.3.4 Collagen crosslinks

Most of the biophysical properties of collagen such as the stability and the tenacity arise through the exceptional folding of many collagen molecules into collagen fibrils and subsequent crosslinking of these fibers. Enzymatic crosslinking occurs at specific lysine or hydroxylysine residues located at the telopeptide region of collagen and links an end of a collagen fibril with an intermediate lysine or hydroxylysine residue in the triple helical region of another collagen fibril. The aldol condensation that links the fibrils occurs spontaneous after extracellular oxidative deamination of the ε-amino group of the lysines to aldehydes by a lysyl oxidase enzyme [85]. The aldehydes cannot only react with each other but can also condensate with another ε-amino group of a lysine, hydroxylysine or a glycosylated hydroxylysine residue [86] resulting in a trivalent crosslink.

Lysyl oxidase expression is regulated, among others, by hypoxia inducible factors [87]. Increased lysyl oxidase expression was found in hypoxic tissues such as tumors [88], where secreted lysyl oxidase leads to tissue remodeling and stiffening of the tumor's surroundings [89]. This finally results in an increased rate of metastasis [90].

1.4 BIOMEDICAL APPLICATION AND PRODUCTION OF RECOMBINANT COLLAGEN

Collagen has a plethora of biomedical applications. The global market size of collagen for regenerative and cosmetic medicine including wound healing and bone graft substitutes in 2014 was estimated at US $ 15 billion [91]. Mostly extracted from animal sources, collagen is applied in wound care management as skin replacement, in reconstructive surgery as bone substitute, in ophthalmology, in drug delivery and for basic matrices in cell culture systems. It is biodegradable and has cell growth stimulating effects. Due to a very high similarity of collagens in mammals, bovine and porcine collagen can be applied in humans with minimal antigenicity. Even though collagens from animal origin are categorized as safe, 1 - 3% of reported cases showed hypersensitivity towards bovine collagen [92]. Human derived collagens from cadavers or placenta are expensive and must be purified from and tested for virus contaminations such as hepatitis or HIV. As an alternative, medical applications use recombinant collagens. However, the existing expression systems are very expensive, cumbersome due to the necessity of collagens posttranslational modifications and are therefore not yet suitable for the large collagen amounts needed in biomedical applications.

1.4.1 BIOMEDICAL APPLICATIONS OF COLLAGEN

Patients losing a substantial fraction of their total body surface, through fire or other accidents, face several serious threats. Every wound represents a threat for bacterial and viral infections that leads to sepsis if not treated appropriately. As intact skin further serves as a barrier for moisture, its damage can lead to massive dehydration and shock. Thus, covering and closing of wounds is of utmost importance. A widely used method to treat open wounds is a dermal regeneration template (DRT) [93]. It consists of a collagen-glycosaminoglycan (GAG) scaffold (98%/2% (w/w)). The collagen scaffold can be engineered precisely in order to have a pore size and a degradation rate suitable for fibroblast, macrophage and neutrophil migration into the wound site and subsequent replacement of the dermal regeneration template by these cells [94]. This results in faster healing, less scarring and immediate physiological wound closure.

Through crosslinking and variation of the collagen to GAG ratio the degradation rate can precisely be defined. This allows incorporation of drugs, such as antibiotics or growth factors, into the DRT matrix in order to be topically released over a prolonged period [95-97]. Antibiotic incorporation in the collagen scaffold results in higher local concentrations of the antibiotic without having detrimental systemic effects as observed when orally administered [98]. Collagen was recently used in combination with lipids for the formulation of drug and gene delivering nanoparticles. Due to the well-characterized pharmacokinetics, the high biocompatibility and its relatively unelaborate production, collagen nanoparticles present promising candidates for future drug delivery systems [99].

Beauty care represents another important market where collagens have become essential. When humans age, collagen type I and III expression in fibroblasts is significantly reduced [100]. This results in wrinkles due to the missing collagens in the connective tissue. Wrinkles can be treated with intradermal injections of collagen. So far, sources for collagen used in dermal fillers comprise bovine, porcine, human cadaver and human fibroblasts from cell culture systems. Even though collagen was the most used dermal filler until 2005, hyaluronic acid based fillers gained marked shares and are now used in more than 50% of all cosmetic applications in the USA [101] due to improved biocompatibility and longevity compared to collagens.

1.4.2 Recombinant production of collagen

Native human collagen is heavily modified upon expression in the endoplasmic reticulum. These posttranslational modifications, most importantly prolyl hydroxylation, are essential to stabilize the collagen triple helix and thereby ensure stability at body temperatures. Hence, collagen expression systems must not only express collagen in large amounts but also precisely modify the protein to meet the requirements of the application.

1.4.2.1 Collagen expression in Bacteria

Classical biotechnological expression systems for recombinant proteins encompass bacteria such as *E. coli*. Bacterial collagens, originated from *Streptococcus pyogenes* could be expressed at high levels exceeding 19 g/L culture [102-104]. These collagens do not require posttranslational modification and self-assemble into a triple helix with T_m of 35 – 39°C. Although from bacterial origin, these collagens are neither immunogenic nor cytotoxic [105], a comprehensive clinical study to evaluate its efficacy and safety in humans is however missing.

Expression of recombinant human collagens in *E. coli* was only of limited success. Since human collagens need prolyl hydroxylation for triple helix stability, the expression system must contain the enzymes for prolyl and lysyl hydroxylation. While unhydroxylated collagen type II was expressed at amounts

exceeding 10 g/l, coexpression with prolyl 4-hydroxylases decreased expression levels remarkably [106]. Human prolyl 4-hydroxylase consists of an α2β2 tetramer in which the α-subunits contain the catalytic domain and the β subunits (protein disulfide isomerases) ensure the solubility of the enzyme. Even though the human prolyl 4-hydroxylase enzyme could be successfully coexpressed with collagen in bacteria, it lacked sufficient activity towards collagenous peptides in order to ensure triple helix formation or, due to the coexpression of several cofactors, only yielded low amounts of the recombinant collagens [107, 108].

Another approach to express hydroxylated human collagen involved an engineered *E. coli* strain with increased prolyl aminoacyl-tRNA synthase. Due to the increased availability of prolyl-tRNA, hydroxyproline was successfully incorporated in α1(I) collagen fragments by addition of hydroxyproline to the hyperosmotic growth medium [109]. Incorporation of hydroxyproline however occurred not only at the Y position of the Gly-X-Y triplets, but also at the X position and results in destabilization of the collagen triple helix.

1.4.2.2 Collagen expression in Yeast

The large majority of studies on heterologous collagen expression were performed in the yeast *Pichia pastoris*. The human prolyl 4-hydroxylase was coexpressed with collagen type I, II and III and contained hydroxylation levels comparable to native human collagens [110]. The recombinant human collagens folded triple-helically but more than 90% of the collagenous proteins accumulated in the ER [111]. Nevertheless, expression levels could be improved by genetic modification of the expression hosts or by modification of the collagens. By removal of the *N*-terminal propeptide, expression levels were increased from 15 mg/L to 20 mg/L. Replacement of the *C*-terminal propeptide with a bacterial trimerization domain and the use of codon-optimized, synthetic genes together with improved fermentation control parameters enhanced the expression level further to 1.6 g/L [112, 113].

Biocompatibility of the recombinant human collagen type III derived of *P. pastoris* was tested in an animal model in order to induce platelet aggregation [114]. The recombinant collagen was able to stop bleeding in a rabbit spleen six times faster than bovine collagen I. There were no adverse detrimental effects from potential contaminants originated from the expression host. Recombinant collagen in form of a collagen sponge has proven to induce less inflammatory response compared to animal derived material. Recombinant human collagen type I is less porous than animal derived collagen and therefore more resistant to bacterial collagenases [115].

Next to *P. pastoris*, recombinant human collagens were also expressed in *Saccharomyces cerevisiae*. By coexpression of the chicken prolyl 4-hydroxylase with human procollagen α1(I) and α2(I), triple helical collagens with hydroxyproline levels of 82% compared to native human collagen type I were accomplished [116]. The collagen consisted however not only of the heterotrimeric α1(I) and α2(I)

collagens but to the same amount also of homotrimeric α1(I) collagen. *S. cerevisiae* derived type III collagen could be triple helically expressed even in absence of both *N*- and *C*-terminal propeptides [117].

1.4.2.3 Collagen expression in N. tabacum

Plants present a cost efficient and easy to manipulate expression system for a variety of small proteins up to large antibodies [118, 119]. Prolyl 4-hydroxylase derived from plants was shown to be active towards collagen but lacked sequence specificity and hydroxylated the X and Y position of Gly-X-Y triplets [120, 121]. Moreover, plants have different N-glycosylation and O-glycosylation patterns that are different to those of humans, which are likely to be immunogenic [122]. To circumvent these obstacles, human procollagen α1(I) and α2(I) cDNAs were coexpressed together with the cDNAs coding the human prolyl 4-hydroxylase and lysyl hydroxylase 3 enzymes. The recombinant collagen was expressed at 1 g/kg dry tobacco leaves and was extracted from the vacuoles. The vacuoles represent a closed environment, thereby preventing undesirable plant specific N-glycosylation. The collagen was assessed for its biocompatibility evaluating binding and proliferation of adult peripheral blood-derived endothelial progenitor-like cells [123].

1.4.2.4 Collagen expression in insect cells

Insect cells, such as *Spodoptera frugiperda* Sf9 cells have become an important expression system for transgenes since the development of the baculovirus-insect cell expression system [124]. Even though Sf9 cells express endogenous prolyl 4-hydroxylase, its activity was too low to stabilize the recombinant collagens. In order to extract stable triple helical recombinant human collagen, the human prolyl 4-hydroxylase was coexpressed together with human collagen type III. The hydroxyproline content of the recombinant human collagen was comparable to native human collagen as assessed by the thermal stability and by amino acid analysis. The yield could be improved from 6 mg/L of culture to 40 mg/L trough expression in high five cells in suspension.

1.4.2.5 Collagen expression in mammalian cells and transgenic animals

Mammalian cell lines harbor a big advantage over all the other expression system regarding collagen production due to sufficient endogenous expression of the prolyl 4-hydroxylase. Depending on the cell line used, recombinant human collagen type I, II, III and V with hydroxylation and glycosylation patterns indistinguishable from native human collagens were produced in amounts ranging from 0.35-2 mg/L in HT1080 up to 15 mg/L in the HEK293 cell line [125-127]. The collagens were secreted in the medium and allowed relatively unelaborate purification.

Next to the expression of human collagens type I, II, III and V, the less common collagen type VII (>0.001% of total extracted collagen from human skin) was expressed in HEK293 and CHO cells for studies focusing on dystrophic epidermolysis bullosa [128, 129]. Recombinant collagen type VII was expressed as triple helical fibril at 2 - 5 mg/L. None of the studies analyzed hydroxylation levels but showed triple helix

stability by chymotrypsin digestion at 37°C for 3 hours. The collagen was intravenously injected in a dystrophic epidermolysis bullosa mouse model and increased life span while reducing clinical symptoms [121]. There were no anti-collagen type VII antibodies expressed in mice upon treatment with recombinant collagen type VII.

Due to the high secretory capability of mammary glands, transgenic mice were engineered carrying the αS1-casein promoter fused to the human *COL1A1* gene. Prolyl- or lysyl hydroxylation has never been described for milk proteins. Nevertheless, homotrimeric collagen $\alpha1(I)_3$ was triple helically secreted in milk of the transgenic mice and showed around 50% of lysyl and prolyl hydroxylation compared to native human collagen. Expression levels reached 8 mg/ml milk [130].

1.4.2.6 Comparison of recombinant human collagen expression systems

In order to produce recombinant human collagen for biomedical applications, expression must be cost efficient and highly reproducible, the collagen needs to be biocompatible and contain the appropriate posttranslational modifications. Ideally, the posttranslational modifications should be tunable to enable a variation of fibril diameter and pore size. This in turn allows selective biodegradability of devices and drugs coupled to the devices. The currently available expression systems for recombinant human collagens are summarized in table II.

While recombinant human collagen expression in bacteria is promising concerning yield and cost efficiency, its lack of endogenous prolyl hydroxylase and inactivity of the human prolyl 4-hydroxylase impeded development of medical devices. Human collagen derived from the yeast strain *P. Pastoris* and from *N. tabacum* too are inexpensive in production and safe in application as shown in several animal models. Expression of recombinant human collagen in mammalian cell lines is possible, but due to expensive growth and purification requirements less lucrative than yeast or plant derived products.

Yeast and plant derived recombinant human collagens are successfully commercialized in two companies. FibroGen Inc. distributes pure recombinant human collagen type I and III for medical applications. Collplant uses recombinant collagen type I in readily available formulations as bone graft matrix, soft tissue repair matrix or wound filler.

Table II: Comparison of the available expression systems for recombinant human collagens.

Expression Host	Col. Expressed	Yield (mg/l)	Advantages	Disadvantages
Bacteria	Bac. Col.	19000	Very high yield inexpensive	Elaborate purification, *In vivo* testing missing
Yeast	Human Pro α1(I, III)	1500	High yield inexpensive	Elaborate purification, coexpression of P4H
Plants	Human Pro α1,2(I)	1 g/kg dry weight	High yield inexpensive	Possible plant glycosylation immunogenic
Mammalian cell culture	Human α1(I,II,III,V, VII)	0.35 - 5	Secreted col., authentic product	Low yield, expensive production
Transgenic mice	Human α1(I)	8000	high yield, authentic posttranslational modifications	Very expensive development costs, low mouse milk production

1.5 BACTERIAL AND VIRAL COLLAGENS

1.5.1 BACTERIAL COLLAGENS

The occurrence of collagenous proteins was previously believed to be restricted to multicellular organisms. With the increased availability of genomic sequencing data of microorganisms, collagen typical $(Gly\text{-}X\text{-}Y)_n$ repeats were identified in genomes of bacteria. Out of 136 analyzed genomes, 25 bacteria contained 56 proteins with collagenous domains ranging from 7 - 745 Gly-X-Y repeats [131]. All of the predicted bacterial collagen-like proteins are flanked by non-collagenous domains and share the structural organization of mammal collagens [102]. Different to human collagens, bacterial collagens occupy positions X and Y less frequently with proline but often incorporate threonine or charged amino acids [131]. This allows triple helical folding of the collagens even in absence of hydroxyproline. Collagens in the deep-sea hydrothermal vent worm *Riftia pachyptila* [132, 133] also incorporate threonine predominantly at the X position. These threonine residues are glycosylated and further stabilize the triple helix. Even though such glycosyltransferases have not yet been identified in bacteria, occurrence of glycosylated collagen in *B. anthracis* suggests their existence [134].

Sequence similarity of bacterial compared to mammal collagens infers triple helicity, but only eight bacterial collagen-like proteins are confirmed to fold triple helically with T_m's of 35°C - 40°C [135],[102]. Unlike in most mammals, where collagens fulfill a structural role in connective tissue, bacterial collagens are not part of cell walls but are attached to it. Many, but not all of the bacteria found to express collagen,

are pathogenic (e.g. *S. pyogenes, B. anthracis, S. enterica*). Well-studied members of bacterial collagen-like proteins are Scl1 and Scl2 that are expressed by *Streptococcus pyogenes.* The non-collagenous V-domain of Scl1 was shown to bind to high-density lipoprotein, low-density lipoprotein, factor H, complement factor H-related protein 1, fibronectin and laminin [136-139]. The collagenous domain further interacts with integrins [140]. Hence, it is believed that the collagenous domain facilitates adherence and internalization into host cells, while the non-collagenous V-domain hinders complement activation and mediates evasion of the immune system [141].

1.5.2 Viral collagens

Many bacteria are infected with bacteriophages that carry additional genetic information and are involved in horizontal gene transfer. Next to virulence factors, many phages encode collagen-like proteins as well [142]. Even though there is no proof yet, these collagens are generally annotated as tail fiber proteins. So far, only very few collagen-like proteins from virophages have been studied.

The protein EPclA from the bacteriophage of the pathogenic *E. coli* strain O157:H7 was recombinantly expressed and biochemically and structurally analyzed [143]. The 37 Gly-X-Y repeat containing collagen domain is interjacent of two non-collagenous domains. The protein is trimeric and its collagenous domain exhibits a T_m of 42°C. Since the non-collagenous domains of this protein resemble other non-collagen like domains in phage tail fiber proteins, they probably serve a similar cause as bacterial collagen-like proteins and enable adherence to their bacterial host [144].

Another studied collagen-like protein is gp12 from the bacteriophage SPP1 [145]. The 6.6 kDa protein contains only eight Gly-X-Y repeats, folds triple helically and exhibits a melting temperature of 37°C to 43°C. GP12 binds to the major capsid protein gp13. Binding of gp13 in turn stabilizes trimeric gp12 and increases its melting temperature by 20°C to 53°C. It is believed that this stabilization ensures the cooperative recruitment of freely available gp12 to the bacteriophage capsid when infected cells lyse.

1.6 COLLAGENS IN NUCLEOCYTOPLASMIC LARGE DNA VIRUSES (MINI-REVIEW)

The following chapter contains a review with the title

Collagens in nucleocytoplasmic large DNA viruses

The manuscript is in preparation and is a collaboration of Stephan Baumann (all *in silico* data), Nina Hochhold (*in vitro* data, not part of the manuscript yet) and Thierry Hennet.

1.6 Collagens in nucleocytoplasmic large DNA viruses

Nucleocytoplasmic large DNA viruses (NCLDVs) were first described in 1983 with the identification of the *Paramecium bursaria* chlorella virus 1 [146]. The size of their genomes ranges from 100 kb up to 2.5 Mb double-stranded DNA, which often encode more than 1'000 genes and thereby exceed genome sizes of numerous parasitic bacteria [147]. All NCLDVs share several orthologous genes such as a DNA polymerase, RNA polymerase and transcription factors [148]. Some of these virus families are packed in giant virus capsules ranging from 700 - 1200 nm in diameter size. These giant viruses are composed of members of the *Mimiviridae, Megaviridae, Pandoraviridae and Pithoviridae*. Even though the international committee on taxonomy of viruses did not yet assign a clear classification to the giant viruses, a certain relation and a common ancestor can be inferred from gene analysis of the conserved viral DNA polymerase [149].

Mimiviridae and *Megaviridae* exhibit a close relationship concerning their DNA polymerases, and might share a common ancestor from the NCLDV family (Fig. 4). *Pandoraviridae* have another origin and cluster more closely together with coccolithoviruses that are part of the *Phycodnaviridae* family. Pithovirus display the closest similarity to *Iridoviridae* and *Marseilleviridae* [150]. All giant viruses show a close relationship based on their DNA polymerases, but reveal otherwise a patchy genome acquired from their hosts, from other viruses and from bacteria [150].

Genomic analysis of the well characterized *Acanthamoeba polyphaga mimivirus* led to the discovery of several for virus uncommon proteins such as collagens, a collagen prolyl 4-hydroxylase and a bifunctional procollagen lysyl hydroxylase and glycosyltransferase [151-153]. Due to the occurrence of collagens in the mimivirus and its close relationship to other giant viruses, we hypothesized collagens being also part of other NCLDVs. We analyzed genomic data from more than 60 members of the NCLDVs (Table S1) and identified 142 collagen-like protein coding genes containing at least 10 uninterrupted repeats of Gly-X-Y (Fig. 4). The occurrence of collage-like genes correlates to the phylogeny as based on the viral DNA polymerases (Fig. 4). We found collagen-like genes in all members of the *Pandoraviridae, Megaviridae, Mimiviridae* and *Pithoviridae,* except in the *Cafeteria roenbergensis* virus and the Samba virus. The giant viruses from these families encode an average of 8.125 (n=16) collagen-like proteins per genome ranging up to 16 in the *Pithovirus sibericum*. We further found two collagen-like genes in the genome of the sputnik virophage, two in the genome of the zamilon virophage and three in the organic lake virophage (Table III). Sputnik and zamilon parasitize the collagen containing mimivirus. In contrast, the organic lake virophage infects the organic lake virus that is deficient in collagen-like genes.

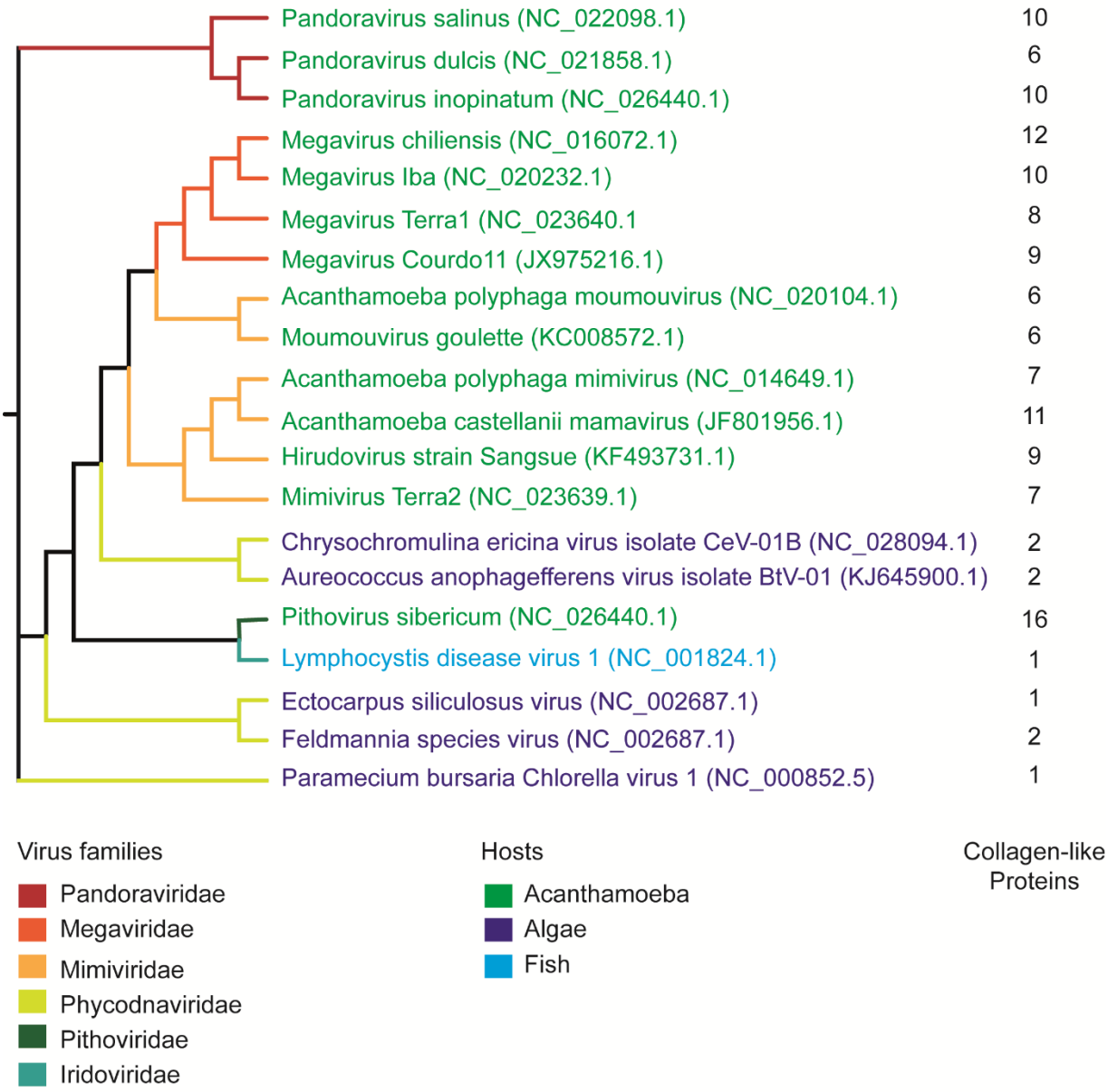


Figure 4: Cladogram of the phylogenetic comparison of the viral DNA polymerase from collagen-like protein containing viruses. Sequence alignment was produced with Clustal-Omega algorithm. The tree was calculated with maximal likelihood using the Mega6 software. Colors of the cladogram branches indicate virus family affiliation, the color of the scientific name indicates host specificity.

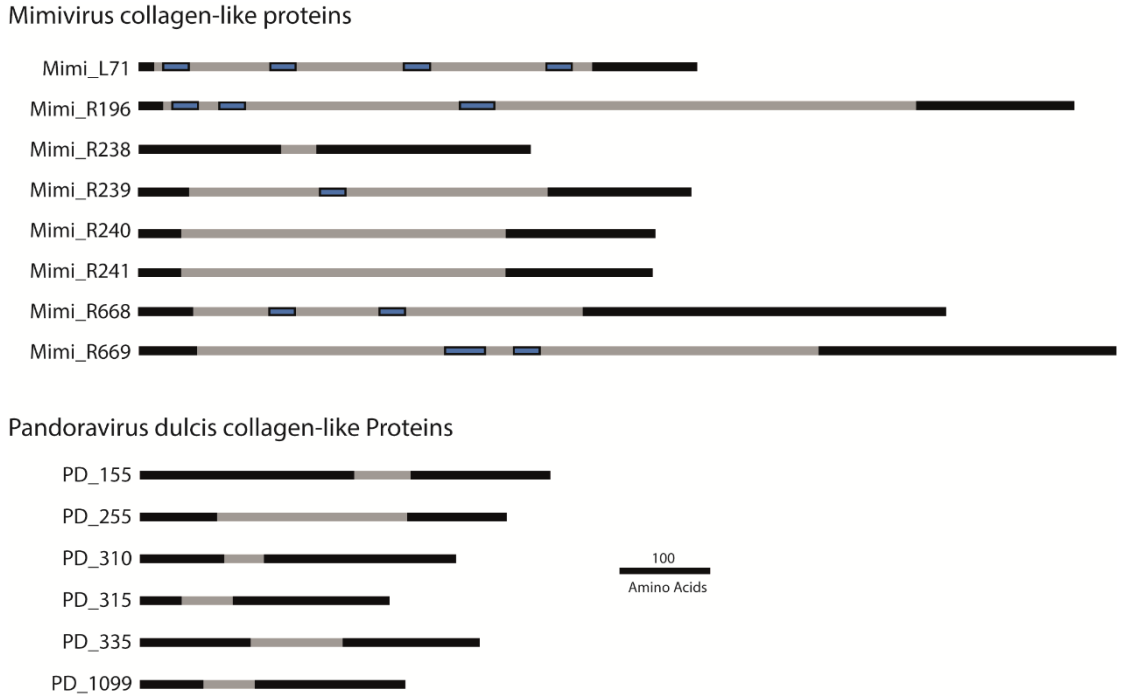
Table III: Virophage species that encode collagen-like proteins

Virus Family	*Scientific Name*	*Genome size / kb*	*Collagen-like genes*	*Host*	*Accession-number*	*Reference*
Phages	*Sputnik virophage*	18.34	2	*Mimivirus/ Acanthamoeba*	NC_011132.1	[154]
	Sputnik 2 virophage	18.34	2	*Mimivirus / Acanthamoeba*	JN603369.1	[155]
	Sputnik 3 virophage	18.34	2	*Mimivirus / Acanthamoeba*	JN603370.1	[156]
	Zamilon	17.28	2	*Mimivirus / Acanthameba*	NC_022990.1	[157]
	Organic Lake virophage	26.42	3	*Algae / organ. Lake virus*	HQ704801.1	[158]

1.6.1 GENE-STRUCTURE ANALYSIS OF VIRAL COLLAGEN-LIKE PROTEINS

For further analysis, we focused on the collagen-like proteins from *Pandoraviridae, Megaviridae, Mimiviridae* and *Pithoviridae* based on the abundance of collagen-like proteins in their genomes. The collagenous domains from these giant viruses contain 13 to 390 Gly-X-Y repeats that are either uninterrupted or contain 44 or 68 amino acid long interdomains between the collagen domains (Fig 5).

A



B

Mimivirus_R239, Protein ID: AAV50512.1

MFGNCQKQNCCTQRNNTGICYSVCPPQTVITVPGNVTCNASRIYVNTGRPNNCFGNDGDLYLDTNTNNLYYKRDGVWLLVGNLSGSSGPSGPQGPKGEKGSNGDKGDKGEIGIQGLKGE
SGADADKGDKGDKGDKGSKGTKGENGDKGNKGDKGDPGIKGSKGEKGSKGDKGEKGNKGDKGDIGLKGESGADADKGDKGDKGSKGDKGNKGDIGLKGESGADADKGDKGDKGSKGNKG
DKGTKGESGLIGTKGDKGSKGDEGIKGDKGEAGTSILEGSGVPSPDLGNNGDLYIDGMTGLLYAKINDEWVPVTSIKGDKGSKGDTGLKGESGADADKGEKGDPGNKGDKGSKGDKGDT
GLKGESGADADKGDKGDKGEKGDKGEKGDKGDKGDNGEKGEKGEKGDNGDKGEKGDNGEKGEKGDNGEKGEKGNNGEKGEKGDNGEKGEKGDNGEKGEKGDNGEKGEKGDNGEKGVKGD
PGSNGSKGTKGEKGEKGDKGINGDKGDKGSKGDTGIKGEAGTAANKGDKGSKGDKGDKGSKGDKGDIGISIKGDKGDKGDKGSKGDKGDIGIKGESGLSIKGDKGDKGGKGDKGDLGSKGDI
GLKGDKGDKGDVGLKGDKGDKGDVGSKGDKGDKGSKGDKGSKGDTGSKGDKGDKGSKGDKGDKGSKGDKGDIGSIGPKGEKGEAGSTNSLYIGSAQLGNAGDFNDEVIPVTGIAYI
TAVGAGGSGYSGTTGGYGGGGAGGAFINYPVYVSSSQTYSAHVGFGGAQVAGNSNKGENTTITIGNLVLLAGGGEGGTATTGGTGGLVSINGTQITAGAPGGTSGNSGVSSIYINPLVI
GGAGGGGGSNGTNAGNGGNYAGFTGGIASPGVNGGGGGGASLFSNGFNGETGAPTTDSGTNYGAGGGGGGNGTQGGNGSLGYVRIDFYSAP

Pithovirus_AZ76_gp240 , Protein ID: YP_009001142.1

MAVRILSYSPQTGERHISQENNLSRNEGGNNQVPINYRGYTGATGPKGDCGRNGCTGPKGNDGKNGKDGCTGPKGNDGKNGKDGCTGATGPKGDCGKDGCTGPTGPKGDCGKDGCTGPK
GADGKNGKDGCTGPKGDQGLIGCTGLKGDTGPKGDCGKDGCTGPKGADGKDGCTGPTGPKGDCGKDGCTGPKGADGKDGCTGPTGPKGDCGKDGCTGPAGERGPMGSTGPGGPGSCVGT
NRIDLRHPMFLKGCGKCPSLVSSCNESGTLVFTNVALGPTDGSSYIPTSGKISVGTTGIYFISLSVSALPHSCEGLSTTVELCSLTGEERKLVIAQLPLQTIHQNPIYASTTALFPLDP
SIPLEVSFSNKSSAEYDMLHSSLVMFRIY

Figure 5: Structure of viral collagen-like proteins. **A)** Graphical representation of the collagen-like proteins from mimivirus and pandoravirus. Black part: non-collagenous domains, grey part: collagenous domains, blue part: interdomains. **B)** Sequence of two viral collagen-like proteins. Top: mimivirus protein L71, representing a collagen with interdomains. Bottom: Pithovirus protein gp240, representing collagen-like protein with uninterrupted collagen domain. Glycine residues are highlighted in red, collagenous domain in light grey, glycine-rich c-terminus in dark grey.

Pithovirus sibericum and the *Pandoraviridae* encode only collagen-like proteins with uninterrupted collagen domains. The *Megaviridae* and *Mimiviridae* express 40 collage-like proteins containing uninterrupted collagen domains and 35 collagens with interdomains. The viral collagen-like proteins structure resembles the structure of mammalian and bacterial collagens. It contains non-collagenous *N*- and *C*- terminal domains and the collagenous domain in between.

1.6.2 Non-collagenous domains

The viral non-collagenous *N*-terminal domains range from 25 to 95 amino acids, the non-collagenous *C*-terminal domain from 175 to 636 amino acids. For most terminal domains in the viral collagen-like proteins, no specific function or structure can be derived from only the amino acid sequence. As determined by a Pfam search [159] for all terminal domains, only one known domain was identified. The protein L71 from mimivirus that shares a high sequence homology with the protein MG751 from *Megavirus chiliensis* and glt00863 from *Moumouvirus goulette* encodes a transmembrane exospore domain in the non-collagenous *C*-terminal domain. This domain occurs in many bacteria from the genus *Bacillus*. In *Bacillus anthracis*, the protein BclB consists of a collagen domain adjacent to a C-terminal transmembrane exospore domain and is probably involved in exosporium assembly [160]. In mimivirus, L71 was as well identified as a surface protein [152] and a similar function to that of BclB can be hypothesized.

In 50% of all collagen-like proteins found in the *Mimiviridae* and *Megaviridae* families, the *C*-terminal non-collagenous domains contained stretches of two to eight repeats of glycine (Fig. 5B). Due to the physical properties of glycine, being small and non-polar, it provides conformational flexibility. In humans, glycine-rich repeats often occur in transmembrane domains in form of GXXXG motif with X being any amino acid [161]. Next to transmembrane proteins, glycine rich repeats occur in structural proteins such as keratins (Fig. 6), where the glycine-rich repeats appear in the head and tail region of the proteins and, similar to collagen propeptides, aid in keratin multimerization [162, 163]. Whether the glycine rich regions serve as anchorage in the mimivirus membrane or serve as nucleation site for collagen trimerization cannot be determined from this data.

Human keratin type II – P35908

MSCQISCKSRGRGGGGGGFRGFSSGSAVVSGGSRRSTSSFSCLSRHGGGGGGFGGGGFGSRSLVGLGGTKSISISVAGGGGGFGAAGGFGGRGGGFGGGSSFG
GGSGFSGGGFGGGGFGGGRFGGFGGPGGVGGLGGPGGFGPGGYPGGIHEVSVNQSLLQPLNVKVDPEIQNVKAQEREQIKTLNNKFASFIDKVRFLEQQNQVL
QTKWELLQQMNVGTRPINLEPIFQGYIDSLKRYLDGLTAERTSQNSELNNMQDLVEDYKKKYEDEINKRTAAENDFVTLKKDVDNAYMIKVELQSKVDLLNQE
IEFLKVLYDAEISQIHQSVTDTNVILSMDNSRNLDLDSIIAEVKAQYEEIAQRSKEEAEALYHSKYEELQVTVGRHGDSLKEIKIEISELNRVIQRLQGEIAH
VKKQCKNVQDAIADAEQRGEHALKDARNKLNDLEEALQQAKEDLARLLRDYQELMNVKLALDVEIATYRKLLEGEECRMSGDLSSNVTVSVTSSTISSNVASK
AAFGGSGGRGSSSGGGYSSGSSSYGSGGRQSGSRGGSGGGGSISGGGYGSGGGSGGRYGSGGGSKGGSISGGGYGSGGGKHSSGGGSRGGSSSGGGYGSGGGG
SSSVKGSSGEAFGSSVTFSFR

Figure 6: Protein sequence of human keratin type II. Glycine rich head and tail regions are highlighted in grey

1.6.3 COLLAGENOUS DOMAINS

The collagenous domains of the viral collagen-like proteins range from an average of 26 G-X-Y repeats in the *Pandoraviridae* up to 400 in the *Mimiviridae*. Based on the amino acid composition (Tab. IV), there are two different families apparent. Pithovirus and the *Pandoraviridae* contain rather short collagenous domains (<60 G-X-Y repeats) and have proline as the most prominent amino acid at the X position. Pithovirus incorporates threonine at the Y position of 50% of Gly-X-Y repeats, similar as found in bacterial collagen-like proteins or in collagen from the deep sea worm *Riftia pachyptila [132, 134]*. *Pandoraviridae* exhibit 20 - 27% proline residues at the Y position, similar to human collagens, where hydroxylated prolines at the Y position are essential for triple helix stabilization. Viral collagen prolyl hydroxylases have been identified in many viral genomes [164] but have not yet been found in *Pandoraviridae* or in the pithovirus. However, prolyl hydroxylases exhibit manifold structures and might be encoded by a not yet identified gene [165]. Alternatively, it might not be necessary to stabilize pithovirus collagens by prolyl hydroxylation, since its habitat is located above the polar circle in a rather cold climate. An assimilation mechanism concerning collagen posttranslational modifications was detected by comparison of cold- and warm-water fish. Depending on their habitats, their collagens exhibit different degrees of prolyl hydroxylation [166]. Whether or not collagen-like proteins from *Pandoraviridae* and *Pithoviridae* are hydroxylated and form stable triple helical structures, cannot be answered at this point.

The second group of viral collagen-like proteins consists of the members of *Mimiviridae* and *Megaviridae*. The collagenous domains are longer compared to pithovirus and *Pandoraviridae* ranging from 33 to 395 G-X-Y repeats. In contrast to mammalian collagens and the collagen-like proteins from pithovirus and *Pandoraviridae*, the *Mimiviridae* collagen-like proteins contain ~ 60% charged amino acids at the X and at the Y position, but barely proline. A similar amino acid frequency is seen in bacteriophages and in bacterial collagen-like proteins [167]. These collagens are not stabilized by hydroxyproline but by electrostatic interactions. The tripeptides GDK or GEK with oppositional charges adjacent to each other were previously studied in the context of triple helix stabilization in host-guest peptides. It was found that the triple helical $(POG)_3$-(XYG-XYG)-$(POG)_3$ polypeptides (O=Hydroxyproline) were destabilized by 10°C by mutation from $(POG)_3$-(PYG-PYG)-$(POG)_3$ to $(POG)_3$-(KYG-KYG)-(POG), but showed an increase in triple helix stability via double mutation to $(POG)_3$-(KDG-KDG)-$(POG)_3$ [168]. The stabilizing effect is a result from ionic interactions between the collagen strands and additionally influences supramolecular assembly in collagen fibrils. In mammalian collagens, charged amino acids make up 20% in average at the X position and 12% in the Y position. The frequency however ranges up to 28% at the X position and 27% in the Y position in collagen type VII. It is believed that the high content of charged amino acids helps stabilizing the collagen domains that are interrupted by non-collagenous domains via electrostatic interactions [169]. The stabilizing effect has only been shown in relatively short bacterial collagenous peptides, but never in large collagen-like proteins as found in *Mimiviridae* and *Megaviridae*. All members

of these two families contain, besides collagen-like genes, also collagen-modifying enzymes. The mimivirus protein L230 is a bifunctional collagen lysyl hydroxylase and collagen hydroxylysyl glucosyltransferase. Modification of the highly abundant lysyine residues by hydroxylation or glucosylation could have an influence on triple helix stability, which will have to be studied experimentally.

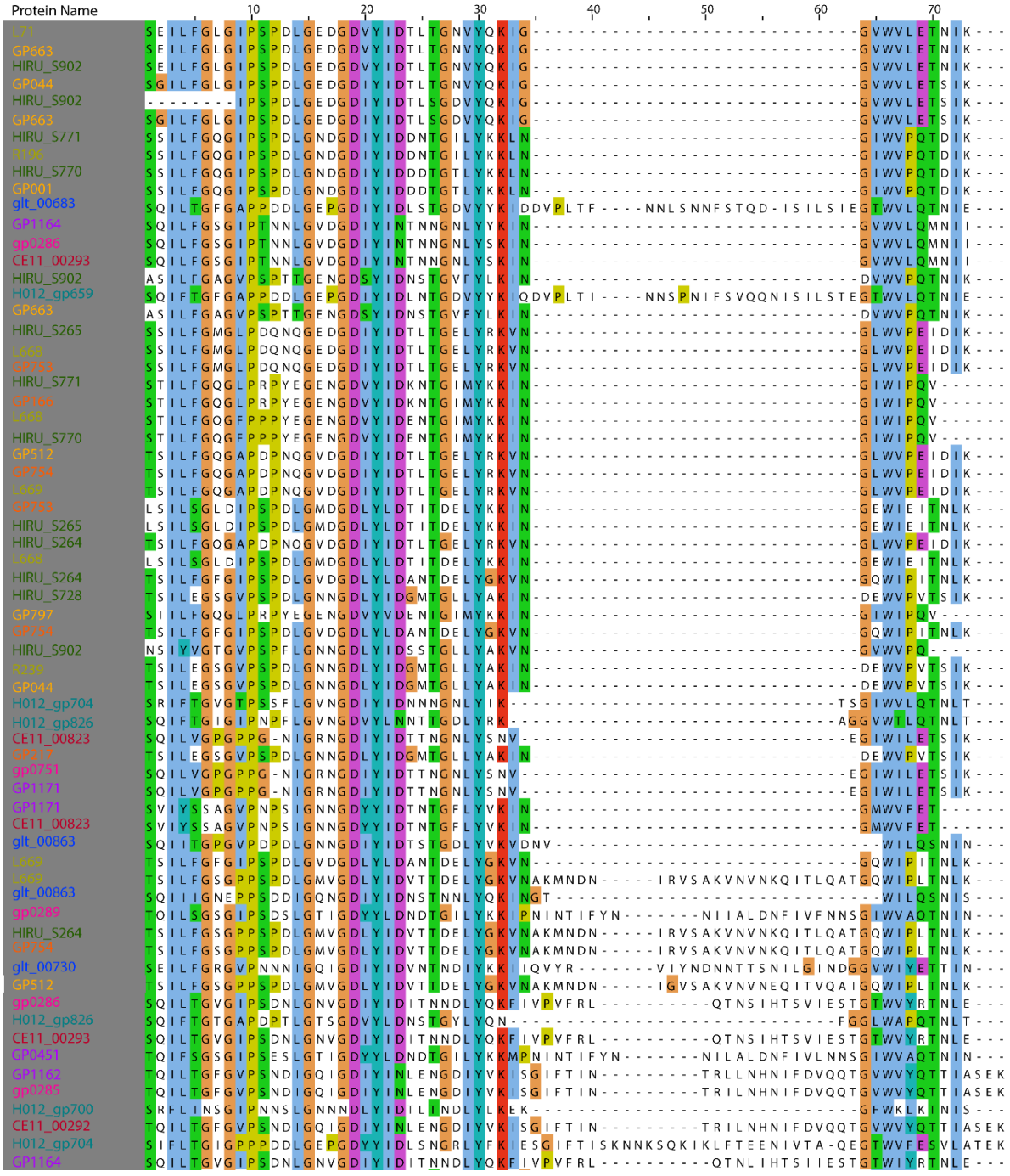
Based on the findings of multimerization domains in the non-collagenous *C*-terminal domain as well as long collagenous domains with interesting features such as a high percentage of charged residues, a triple helical conformation of viral collagen-like proteins can be hypothesized, but still needs to be proven experimentally.

		Three most occurring AA's at pos. X			Three most occurring AA's at pos. Y		
Name	Ø G-X-Y-repeats/ Protein	1.	2.	3.	1.	2.	3.
Mimivirus	238	49% D	11% E	9% S	64% K	7% I	6% N
Mimi. Terra 2	201	46% D	13% E	10% S	62% K	7% I	5% N
Mimi. Isolate 4	246	45% D	12% E	10% S	65% K	7% I	6% N
Megav. Chili.	122	43% D	18% E	9% L	49% K	9% I	7% L
Megav. Courdo	116	41% D	20% E	10% N	49% K	9% I	8% L
Megav. Terra 1	136	45% D	18% E	10% L	50% K	10% I	6% L
Moumouvirus	168	52% D	12% E	6% I	53% K	16% L	5% G
Moumou. Goul.	147	55% D	12% E	7% I	59% K	16% L	5% G
Hirudovirus	217	46% D	12% E	7% I	64% K	6% I	6% N
Mamavirus	159	45% D	12% E	8% I	62% K	7% I	6% N
Megavirus Iba	112	40% D	17% E	10% N	47% K	10% I	8% L
Pithovirus	52	34% P	16% A	6% V	50% T	12% I	11% D
Pandora. Inop.	23	55% P	10% A	5% E	27% P	12% A	10% K
Pandora. Sal.	27	53% P	12% A	5% E	20% P	11% Q	10% A
Pandora. Dul.	29	58% P	8% A	7% E	23% P	16% K	12% Q
Sputnik	26	21% D	15% L	11% E	41% K	13% D	13% T
Zamilon	25	36% D	25% L	8% E	45% K	13% N	8% I
Human Col. 1-10	266	29% P	13% E	9% L	39% P	8% A	6% Q
Bacterial	66	30% E	28% P	12% D	22% Q	20% K	18% R

Table IV: Amino acid composition of the collagenous domains of viral collagen-like proteins. Negatively charged amino acids are depicted in blue, positively charged amino acids in red. Data is evaluated from all viral collagen-like proteins in a given virus, from human collagens Col1α(I-X) or from 10 different bacterial collagen-like proteins from *Staphylococcus* and *Streptococcus* species.

1.6.4 Evolutionary aspects of viral collagen-like proteins

There is an ongoing discussion about the evolution of giant viruses and their taxonomy. Phylogenetic analysis of several conserved genes such as DNA polymerases, aminoacyl-tRNA synthetases and serine/threonine protein kinases showed three distinct origins for the giant virus families. *Mimiviridae* and *Megaviridae* are distantly related to *Phycodnaviridae*, *Pandoraviridae* are related to *Coccolithoviruses* and the Pithovirus is most closely related to *Iridoviridae* and *Marseilleviridae* [150]. Analysis of 128 mimivirus proteins revealed various origins uniformly distributed over *Eucarya* and *Bacteria* [170]. Since horizontal gene transfer was involved in the acquisition of many genes of the mimivirus, this could also apply for viral collagen-like proteins. Phylogenetic analysis of collagenous domains is not feasible due to the conserved G-X-Y repeats. Instead, we identified conserved



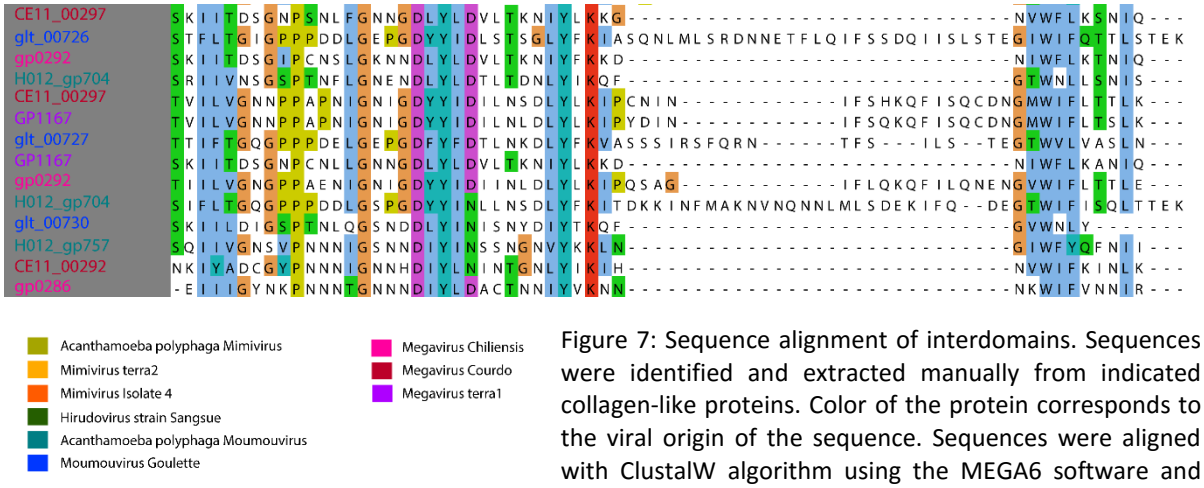


Figure 7: Sequence alignment of interdomains. Sequences were identified and extracted manually from indicated collagen-like proteins. Color of the protein corresponds to the viral origin of the sequence. Sequences were aligned with ClustalW algorithm using the MEGA6 software and standard parameters.

interruptions (interdomains) in the collagenous domains of *Megaviridae* and *Mimiviridae*, which can be used for phylogenetic analysis (Fig. 7 - 9).

The interdomains from the viral collagen-like proteins show a high amount of conserved amino acids. Several amino acids are 100% conserved even though the collagen-like proteins vary in size of the collagenous domains (Fig. 7). There is no relation between collagen length and amount of interdomains. Mimivirus R196 for example contains one interdomain, dividing the 395 GXY repeats long collagenous domain in 155 and 240 GXY repeats. The shorter mimivirus collagen-like protein L71 contains three interdomains, resulting in 10 to 68 GXY repeat containing fragments.

A basic local alignment search for similarity of the conserved interdomain sequence (SSILFGSGIPSPDLGNNGDIYIDTLTGDLYKKINGVWVPQTNIK) (Fig. 8) against the non-redundant protein sequences database from NCBI revealed 14 significant hits (E-value < 10^{-4}) from bacterial origin. Interestingly, all of the found interdomains are located adjacent to collagenous domains and are part of surface proteins (Fig. 9).

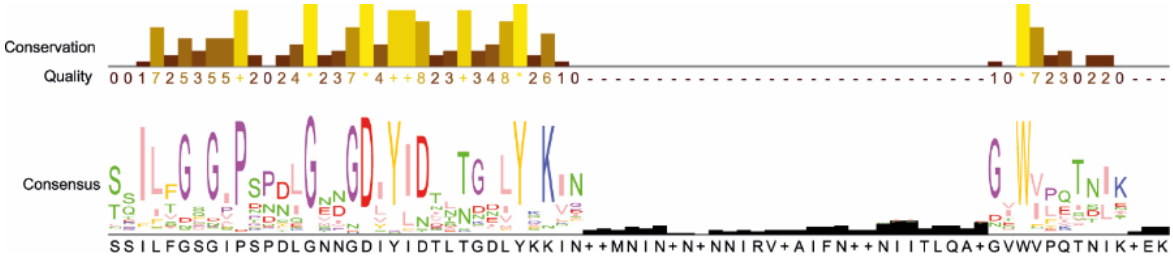
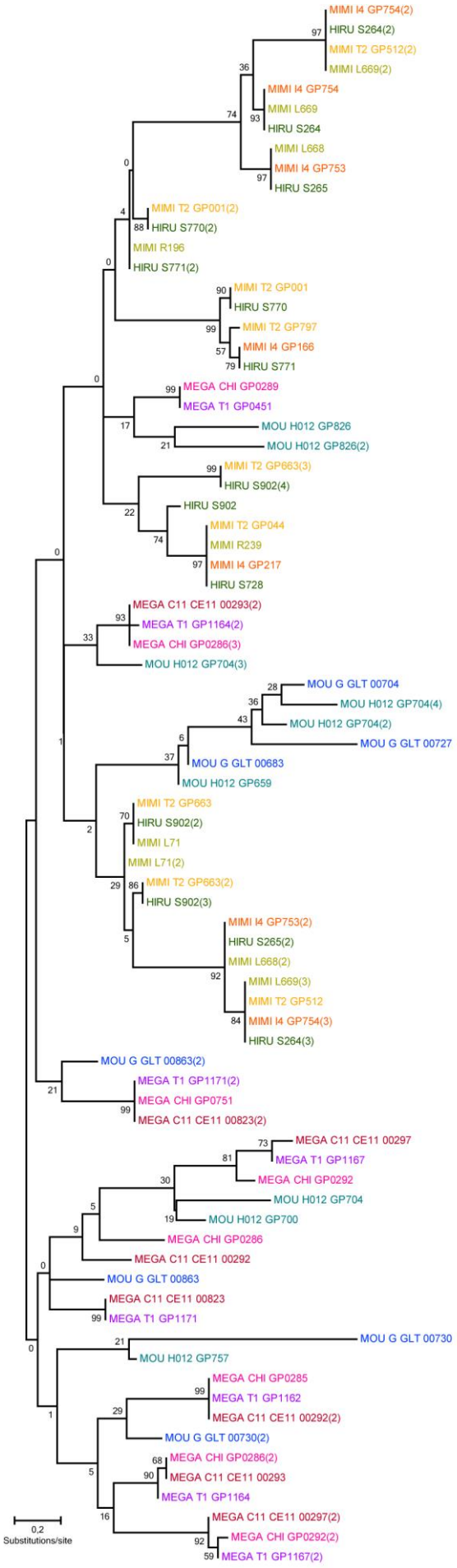


Figure 8: Consensus sequence extracted from sequence alignment of interdomains (Fig. 7). Conservation quality from 1 – 10 based on biophysical properties of conserved amino acids (polarity, aromaticity, size).

YSIRK signal domain/LPXTG anchor domain surface protein [Staphylococcus lugdunensis], WP_050786884

…ADGKNGQNGRNGRQGTELISGITPPTDSDGLNGDTYIDTNTGDTYKKENGKWVPSGNIKGPKGDKGENGQNGRDGKDVLNGKVDPQSSQGKDGDK
YINTVTGDVFVKSGDNWTNEGNIKGPKGDKGDNGLNGRDGKDVLNGKVNPQPSQGKDGDKYINTVTGDVFVKSGGNWTNEGNIKGPKGDRGERGET
GAKGDKGDNGQNGRDGKDILNGKVNPQPSQGKDGDKYINTVTGDVFVKSGGNWTNEGNIKGPKGDRGERGETGPKGDKGENGTSIGIT……

Figure 9: Protein sequence extract from one significant hit identified by BLAST-P search for similarity to the consensus interdomain identified in Fig. 5. Glycines are depicted in red, collagenous domains are highlighted in grey, the interdomains in yellow.



The bacterial proteins identified by the BLAST search contain shorter collagenous domains and resemble the total structure of the viral collagen-like proteins only poorly.

The phylogenetic tree based on the interdomains (Fig. 10) can identify the relation of viral collagen-like proteins between each other. The tree contains two major branches. The top branch predominantly consists of collagen-like proteins from *Mimiviridae*, while the bottom branch contains the *Megaviridae* collagen-like proteins. In between, there are sub-branches containing interdomains from moumouvirus which cluster partially with *Mimiviridae* and partially with *Megaviridae*. Many viral collagen-like proteins contain more than one interdomain. These interdomains cluster closely together (e.g. Mimi_L71 (1) and Mimi_L71 (2), Mimi_L669 (1) and Mimi_L669 (2)) and cluster with the interdomains from the closest homologues from other viruses (Mimivirus, Hirudovirus, Mimivirus Terra2). The interdomains do not only cluster with interdomains from the same proteins, but also with proteins found in close genomic proximity of them (e.g. Mimi_L668 and Mimi_L669 or Mega_Chi_GP0285 and Mega_Chi_GP0286). Duplication of genes is a common feature in giant viruses [171]. It can therefore be assumed, that the variety of collagen-like proteins was developed through gene duplication and subsequent trimming. The close homology of the viral collagen-like proteins infers a common ancestor for the *Mimiviridae* and *Megaviridae* collagens-like proteins, which is not identified until today.

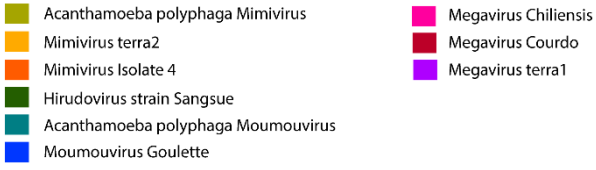


Figure 10: Phylogenetic tree based on protein sequence alignment of interdomains from indicated viral collagen-like proteins. The tree was calculated with maximal likelihood and tested with 500 repeats of bootstrap replications using the Mega6 software

1.6.5 Horizontal gene transfer

Horizontal gene transfer in bacteria can be achieved by vectors such as bacteriophages or plasmids. Several of the giant viruses can be infected by virophages, a phenomenon not observed in other viruses [172]. Similar to bacteriophages, virophages hijack the host-virus during replication in virus factories and use the host-virus' replication machinery for self-proliferation [154]. Among the 21 – 26 genes expressed by these virophages, we identified two to three collagen-like genes (table III). While Sputnik 1 - 3 and Zamilon infect the collagen-like protein containing mimivirus, the organic lake virophage infects the organic lake virus that does not encode any collagen-like genes.

Sputnik virophage protein gp06 and its homolog X812_gp15 from the Zamilon virophage share high sequence similarity with mimivirus and mamavirus protein R196 at the *C*-terminus of the collagen-like protein (Fig. 11). Due to the much longer collagen domain in the viral proteins compared to the virophage proteins, it is unlikely that the viral collagen-like proteins were acquired via the virophages but rather that the virophage collagens originated from its host either partially, or completely with subsequent removal of parts of the collagenous domains as also suggested for other non-collagenous proteins [154].

```
Mimi_R196         KGDKGDKG---IKGDKGDKGDKDDKGNKGDKGDKGDKGIKGDKGD---------KGSEGP
Mama_R196         KGDKGTKGDKGDKGEPGINGLKGDKGNKGNVGDKGDPGINGLKGETGSQGIKGNPGINGL
Sputnik_GP06      KGDKGETGSNGLKGETGSNGLKGETG-SGDKGDKGDSGTDGLKGQAGTDGLKGQAGTDGL
                  ***** .*    **: * :* * :.* .*: ***** * .* **:          * :*

Mimi_R196         KGDKGDQGINGLKGDKGDKGEPGTNGLKGDKGDPGTNGLKGETG------SQGIKGDPGT
Mama_R196         KGETGSQGIKGNPGINGLKGETGNQGIK---GNPGINGLKGETG------NQGIKGDPGT
Sputnik_GP06      KGDSGTDGLKGQTGNDGLKGQAGNDGIKGDKGEPGNDGLKGQTGTQGIKGDPGAKGDPGT
                  **:.* :*::*  * .* **: *.:*:*   *:** :****:**      . * ******

Mimi_R196         SVVVSSGIWTPSVTFSTIFTGITPSTSIYSRIGNIVTFSIFTNATIPANMASGIIYFTLP
Mama_R196         SVVISSGIWTPSVTFSTIFTGITPSTSIYSRIGNIVTFSIFTNATIPANMASGIIYFTLP
Sputnik_GP06      SVVLSSGIWTPSVTFSTIFTGITPSTSIYSRIGNIVTFSIFTNATIPANMASGIIYFTVP
                  ***:**************************************************:*

Mimi_R196         VPSGDFNGNNVIGTASITGTSTGVVQYGILTASASTPAIILSMKIQASSSISVAIYATGQ
Mama_R196         VPSGDFNGNNVIGTASITGTSTGVVQYGILTASASTPAIILSMKIQASSSISVAIYATGQ
Sputnik_GP06      VASSDFNGNNVIGTASITGNSTGVVQYGILTATASTAGIVLSMKIQASTSISVAIYATGQ
                  * *.***************.************:*** .*:********:**********

Mimi_R196         YFIPPG
Mama_R196         YFIPPG
Sputnik_GP06      YFIPPG
                  ******
```

Figure 11: Sequence alignment of the C-terminal sequence from mimivirus and mamavirus proteins R196 and the Sputnik protein GP06.

1.6.6 Conclusions

We identified 142 collagens in the genomes of 60 analyzed NCLDVs. By analyzing the amino acid composition and the structure of the collagen-like genes, we found different phylogenetic origins of the viral collagens. *Pandoraviridae* and pithovirus contain rather short collagenous domains (15-60 Gly-X-Y repeats) with 34 - 58% proline at the X position. While *Pandoraviridae* occupy the Y position predominantly with 20 – 27% proline, pithovirus incorporates threonine in 50% of all Gly-X-Y repeats. In

contrast, collagen-like proteins from *Mimiviridae* and *Megaviridae* have longer (33 – 395 Gly-X-Y repeats) collagenous domains that are interrupted with conserved interdomains in 50% of the collagen-like proteins. The collagens from *Mimiviridae* and *Megaviridae* further incorporate 47 – 64% oppositely charged amino acids at the X and the Y position of Gly-X-Y repeats. A phylogenetic analysis of the interdomains revealed a close phylogenetic relationship between the collagen-like proteins from *Mimiviridae* and *Megaviridae*. The phylogenetic analysis further indicates gene duplication for some of the collagens found in *Mimiviridae* and *Megaviridae*. Mimivirus isolate 4, which was obtained by 150 passage cycles of the original *Acanthamoeba polyphaga* mimivirus contained interdomains identical to the interdomains seen in the original mimivirus [173]. This indicates that mutations are not commonly incorporated but need at least several hundred passages to occur. While a close relationship for collagen-like proteins between *Mimiviridae* and *Megaviridae* is proven, more data needs to be collected in order to produce a reliable phylogenetic analysis also for *Pithoviridae* and *Pandoraviridae*.

Based on the collagen-like amino acid composition, triple helical folding of the collagen-like proteins is probable but experimental proof is missing. It can be assumed that collagen-like proteins from *Mimiviridae* and *Megaviridae* use a mechanism for triple helix stabilization involving electrostatic interactions of the collagen molecules. Pandoraviridae might use prolyl hydroxylation for helix stabilization, even though such enzymes have not yet been identified in the genomes of these viruses. Hydroxylation of prolines stabilizes collagens if incorporated at the Y position but hardly have an effect on collagen stability if incorporated at the X position [174, 175]. Pithovirus incorporates threonine at 50% at the Y position of all Gly-X-Y repeats resulting in a total content of 8 – 21% threonine in the collagen-like proteins. A similar pattern of threonine is used in collagens from the deep-sea worm *Ryftia pachyptila* that uses > 16% threonine in order to stabilize collagen. In *Ryftia pachyptila*, it was shown that glycosylation on threonine residues can stabilize the collagen up to a melting temperature of 37°C [132]. Further experiments will show whether pithovirus uses a similar strategy to stabilize the collagen-like proteins.

Table S1: Genomes searched for collagen-like proteins

Name of the virus	Accession number
Asfarviridae	
African swine fever virus strain Ken05/Tk1	KM111294.1
African swine fever virus E75	FN557520.1
African swine fever virus strain BA71V	NC_001659.2
Ascoviridae	
Heliothis virescens ascovirus 3e	NC_009233.1
Trichoplusia ni ascovirus 2c	NC_008518.1
Spodoptera frugiperda ascovirus 1a	NC_008361.1
Iridoviridae	
Aedes taeniorhynchus iridescent virus	NC_008187.1
Invertebrate iridovirus 22	NC_021901.1
Invertebrate iridovirus 25	NC_023613.1
Invertebrate iridescent virus 6	NC_003038.1
Lymphocystis disease virus - isolate China	NC_005902.1
Lymphocystis disease virus 1	NC_001824.1
Turbot reddish body iridovirus	GQ273492.1
Rock bream iridovirus	KC244182.1
Infectious spleen and kidney necrosis virus	NC_003494.1
Orange spotted grouper iridovirus	AY894343.1
Rock bream iridovirus strain RBIV-KOR-TY1	AY532606.1
Large yellow croaker iridovirus	AY779031.1
Red sea bream iridovirus	AB104413.1
Andrias davidianus ranavirus isolate 2010SX	KF033124.1
Andrias davidianus ranavirus isolate 1201	KC865735.1
Common midwife toad ranavirus	JQ231222.1
Tiger frog virus	AF389451.1
European sheatfish virus	NC_017940.1
Ambystoma tigrinum virus	KR075886.1
Singapore grouper iridovirus	NC_006549.1
Chinese giant salamander iridovirus	KF512820.1
European sheatfish virus	NC_017940.1
Ambystoma tigrinum stebbensi virus	AY150217.1
Epizootic haematopoietic necrosis virus	NC_028461.1
Singapore grouper iridovirus	NC_006549.1
Soft-shelled turtle iridovirus	EU627010.1
Marseilleviridae	
Cannes 8 virus	KF261120.1
Tunisvirus fontaine2 strain U484	KF483846.1
Marseillevirus	NC_013756.1
Lausannevirus	NC_015326.1
Lausannevirus isolate 7715	HQ113105.1

Megaviridae	
Megavirus Chiliensis	NC_016072.1
Megavirus courdo11	JX975216.1
Megavirus lba isolate LBA111	NC_020232.1
Megavirus terra1	NC_023640.1
Mimiviridae	
Sambavirus	KF959826.1
Hirudovirus strain Sangsue	KF493731.1
Moumouvirus goulette	KC008572.1
Acanthamoeba polyphaga moumouvirus	NC_020104.1
Acanthamoeba polyphaga mimivirus	AY653733.1
Acanthamoeba polyphaga mimivirus isolate 4	JN036606.1
Mimivirus terra2	NC_023639.1
Cafeteria roenbergensis virus	NC_014637.1
Acanthamoeba castellanii mamavirus	JF801956.1
Molliviridae	
Mollivirus sibericum isolate P1084-T	NC_027867.1
Pandoraviridae	
Pandoravirus Salinus	NC_022098.1
Pandoravirus Dulcis	NC_021858.1
Pandoravirus inopinatum	NC_026440.1
Phycodnaviridae	
Paramecium bursaria Chlorella virus 1	NC_000852.5
Phaeocystis globosa virus strain 16T	NC_021312.1
Ectocarpus siliculosus virus 1	NC_002687.1
Ostreococcus lucimarinus virus OlV1	NC_014766.1
Ostreococcus tauri virus 1	NC_013288.1
Yellowstone lake phycodnavirus 1	NC_028112.1
Micromonas sp. RCC1109 virus MpV1	NC_014767.1
Chrysochromulina ericina virus	NC_028094.1
Aureococcus anophagefferens virus	KJ645900.1
Feldmannia species virus	NC_011183.1
Pithoviridae	
Pithovirus sibericum isolate P1084-T	NC_023423.1
Poxviridae	
Vaccinia virus	NC_006998.1
Parapoxvirus red deer/HL953 strain HL953	NC_025963.1
Canarypox virus	NC_005309.1

2. RESULTS

This thesis resulted in two submitted manuscripts.

Recombinant expression of hydroxylated human collagen in *Escherichia Coli*

The manuscript was published in in Journal of applied microbiology and biotechnology in 2013 and was done in collaboration with Christoph Rutschmann, Jürg Cabalzar, Kelvin B. Luther and Thierry Hennet.

Own Contribution:

Planning and coordination of the study

Cloning and expression of recombinant human collagen type III

Coexpression of recombinant human collagen type III with viral prolyl 4-hydroxylase and lysyl hydroxylase

Analysis by circular dichroism

Writing of the introduction

The second manuscript with the title

Collagen accumulation in osteosarcoma cells lacking GLT25D1 collagen galactosyltransferase

was submitted in February 2016 to the Journal of Cell Science. The manuscript was a collaboration with Thierry Hennet. All experiments were performed by myself.

2.1 RECOMBINANT EXPRESSION OF HYDROXYLATED HUMAN COLLAGEN IN *ESCHERICHIA COLI*

Christoph Rutschmann[#], Stephan Baumann[#], Jürg Cabalzar, Kelvin B. Luther,

and Thierry Hennet[1]

Institute of Physiology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich,

Switzerland

[#]both authors contributed equally

[1]To whom correspondence should be addressed: Thierry Hennet, Institute of Physiology, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland, Tel: +41 44 635 5080; Fax: +41 44 635 6814; E-mail: thennet@access.uzh.ch

2.1.1 Abstract

Collagen is the most abundant protein in the human body and thereby a structural protein of considerable biotechnological interest. The complex maturation process of collagen, including essential post-translational modifications such as prolyl and lysyl hydroxylation, has precluded large-scale production of recombinant collagen featuring the biophysical properties of endogenous collagen. The characterization of new prolyl and lysyl hydroxylase genes encoded by the giant virus mimivirus reveals a method for production of hydroxylated collagen. The coexpression of a human collagen type III construct together with mimivirus prolyl- and lysyl hydroxylases in *Escherichia coli* yielded up to 90 mg of hydroxylated collagen per liter culture. The respective levels of prolyl and lysyl hydroxylation reaching 25% and 26% were similar to the hydroxylation levels of native human collagen type III. The distribution of hydroxyproline and hydroxylysine along recombinant collagen was also similar to that of native collagen as determined by mass spectrometric analysis of tryptic peptides. The triple helix signature of recombinant hydroxylated collagen was confirmed by circular dichroism, which also showed that hydroxylation increased the thermal stability of the recombinant collagen construct. Recombinant hydroxylated collagen produced in *Escherichia coli* supported the growth of human umbilical endothelial cells, underlining the biocompatibility of the recombinant protein as extracellular matrix. The high yield of recombinant protein expression and the extensive level of prolyl and lysyl hydroxylation achieved indicate that recombinant hydroxylated collagen can be produced at large scale for biomaterials engineering in the context of biomedical applications.

Keywords: Protein engineering, post translational modification, hydroxylysine, hydroxyproline, virus

2.1.2 Introduction

The structural and functional versatility of collagen in vertebrates makes it a coveted protein for tissue and biomaterials engineering. Yet, the considerable size of collagen polypeptides and the requirement for post-translational modifications have impeded the large-scale production of recombinant collagen featuring the biophysical properties of natural collagen. All types of collagen feature a triple helical conformation composed of repeats of the G-x-y motif, in which proline and lysine often occur at the x and y positions. During translation in the endoplasmic reticulum, selected proline and lysine residues are hydroxylated by dedicated hydroxylases, thereby yielding hydroxyproline (Hyp) and hydroxylysine (Hyl) (Myllyharju and Kivirikko 2004). The formation of Hyp is essential to stabilize the collagen triple helix and confer its thermal stability at body temperature (Shoulders and Raines 2009). Lysyl hydroxylation is involved in the formation of covalent intra- and inter-molecular crosslinks, contributing to condensation and fibril formation (Takaluoma et al. 2007). Hyl also serves as acceptor for the attachment of collagen-specific glycans (Schegg et al. 2009). Defects of lysyl hydroxylation lead to diseases such as Ehlers-Danlos type-VI (Hyland et al. 1992), Bruck syndrome (van der Slot et al. 2003), and skeletal dysplasia (Salo et al. 2008), demonstrating the biological importance of this post-translational modification.

The multimeric organization and limited stability of animal prolyl 4-hydroxylases and lysyl hydroxylases make them poor choices for the efficient production of recombinant collagen in conventional protein expression systems such as bacteria and yeasts, which lack endogenous prolyl and lysyl hydroxylases. Human collagen prolyl 4-hydroxylase has been expressed in *Escherichia coli* (Neubauer et al. 2005; Pinkas et al. 2011), although with limited activity towards short collagenous substrates. The coexpression of human prolyl 4-hydroxylase subunits and collagen constructs in the yeast *Pichia pastoris* has

enabled the production of prolyl hydroxylated collagen up to 1.5 g per liter of culture (Nokelainen et al. 2001). Dual hydroxylation of proline and lysine has not yet been achieved in *Pichia pastoris*. Coexpression of human prolyl 4-hydroxylase subunits and the lysyl hydroxylase LH3 has been described in tobacco plants, in which recombinant human collagen type I was expressed at up to 200 mg per kg of fresh leaves (Stein et al. 2009). The use of animal cells, such as Sf9 insect cells (Lamberg et al. 1996; Tomita et al. 1995) and HEK293 human cells (Fichard et al. 1997) that express prolyl and lysyl hydroxylase endogenously, yields recombinant collagen in the μg to mg range per liter of culture, thereby precluding large-scale applications of the recovered collagen product.

The description of several aquatic giant viruses belonging to *Phycodnaviridae* (Van Etten 2003) and *Mimiviridae* (Raoult et al. 2004) has shown that collagen-like genes are not restricted to metazoans and some prokaryotes. In addition to collagen genes, these viruses harbor genes encoding prolyl 4-hydroxylase (Eriksson et al. 1999) and lysyl hydroxylase enzymes (Luther et al. 2011). These viral hydroxylases are soluble and active when expressed in *E. coli*, thus opening new possibilities for the production of recombinant hydroxylated collagen in bacterial expression systems. So far, human collagen type II has been produced at amounts exceeding 10 g per liter (Guo et al. 2010), although without post-translational modifications. To circumvent this limitation, we now exploit bacterially active prolyl and lysyl hydroxylase enzymes from the giant virus mimivirus (Luther et al. 2011) to produce recombinant hydroxylated collagen at high yield in *E. coli*.

2.1.3 Materials and Methods

Cloning of mimivirus hydroxylase expression vectors - The mimivirus lysyl hydroxylase L230 (Luther et al. 2011) and prolyl 4-hydroxylase L593 open reading frames were amplified by PCR from mimivirus genomic DNA using primers including *Xho*I and *Bam*HI sites. The primers were 5'-TGACCTCGAGATTAGTAGAACTTATGTAATT-3' and 5'-CAGGGATCCGTCCAATAAAGTGTATCAAC-3' for L230, 5'-TGACCTCGAGAAAACTG-TGACTATCATTACAATA-3' and 5'-CAGGGATCCATTTTGTGTTAAAAAAATTTTAGG-3' for L593. The resulting amplicons were ligated as *Xho*I-*Bam*HI fragments into the *Xho*I-*Bam*HI linearized expression vector pET16b, yielding the pET16b-L230 and pET16b-L593 vectors. Expression vectors lacking His-tags were prepared by first amplifying the L230 and L593 genes using the primers 5'GTCGACGAGCTCACCATGGGCATTAGTAGAAC-3' and 5'-GTAATGACATATGCGCAAGCCCAG-3' for L230, 5'-ATACCATGGTATTGTCA-AAATCTTGTGTGT-3' and 5'-CAGGGATCCATTTTGTGTTAAAAAAATTTTAGG-3' for L593. The corresponding amplicons were introduced into pET16b linearized with *Nco*I-*Nde*I for L230 and with *Nco*I-*Bam*HI for L593, yielding pET16b-noH-L230 and pET16b-noH-L593. The bicistronic vector pET16b-L593/L230 was prepared by inserting the expression cassette of the pET16b-L593 as a *Bgl*II-*Hind*III fragment into the *Bam*HI-*Hind*III-linearized pET16b-L230 vector. The bicistronic vector pET16b-noH-L593/L230 featuring L230 and L593 without His-tag was prepared in the same way.

Cloning of collagen expression vectors - A fragment of human collagen type III *COL3A1* cDNA encompassing 1206 bp and lacking propeptide-encoding regions was custom synthesized (GenScript, Piscataway, NJ, USA) using codons optimized for bacterial expression and including *Nco*I and *Bam*HI sites at 5'- and 3'-ends (Fig. 1). The pET28a expression vector was first digested with *Nco*I-*Bam*HI, which eliminates the His-tag at the

N-terminal site. The resulting hCOL3 segment was inserted as a *Nco*I-*Bam*HI fragment into pET28a, yielding pET28a-hCOL3-His.

Protein expression in E. coli - The pET16b- and pET28a-based vectors were transformed into chemically competent *E. coli* BL21 (DE3) cells, which were plated on LB-agar plates containing 50 μg/ml kanamycin (Sigma-Aldrich) and 100 μg/ml ampicillin (Sigma–Aldrich) and incubated overnight at 37°C. Protein expression followed standard protocols (Tolia and Joshua-Tor 2006). Briefly, bacteria were grown in liquid cultures at 37°C under agitation at 220 rpm until reaching an OD_{600} value of 0.6. Isopropyl β-D-1-thiogalactopyranoside (Sigma-Aldrich) was added to 1 mM to induce expression and the cultures were incubated for a further 3 h at 34°C under agitation at 220 rpm.

Protein purification - Cells were pelleted at 4,000 x *g* at 4°C for 30 min, resuspended in 3 ml of 20 mM sodium phosphate, pH 7.4, 100 mM NaCl per gram of *E. coli* wet weight and lysed with 250 μg/ml lysozyme, 4 mg/ml deoxycholic acid under rotation at 4°C for 20 min. DNAse I (Fluka, Buchs, Switzerland) was added to 20 μg/ml and incubation proceeded at room temperature for 30 min. Cell lysates were clarified by centrifugation at 12,000 x *g* for 30 min at 4°C and filtered through 0.22 μm membrane filters (Milipore). Imidazole was added to a concentration of 20 mM and His-tagged proteins were purified by affinity chromatography on a 1 ml HisTrap FF Ni Sepharose 6 column (GE - Healthcare) using an Äkta FPLC system (GE - Healthcare). Elution of His-tagged proteins was performed with 500 mM imidazole, 20 mM sodium phosphate, pH 7.4, 100 mM NaCl. His-tagged proteins were detected after transfer to nitrocellulose (Highbound ECL, GE-Healthcare) using the anti-polyHis HIS-1 monoclonal antibody (Sigma-Aldrich).

Prolyl and lysyl hydroxylase activity assays - Prolyl and lysyl hydroxylase activities were measured as described previously (Luther et al. 2011). Briefly, 5 μg His-tag purified L230 lysyl hydroxylase or L593 prolyl hydroxylase were added to acceptor peptides at 0.5

mg/ml in 50 mM Tris-HCl, pH 7.4, 100 µM $FeSO_4$, 1 mM ascorbate, 100 µM DTT, 60 µM 2-oxoglutarate and 100 nCi of 2-oxo[^{14}C]glutarate (PerkinElmer Life Sciences) in a total volume of 100 µl and incubated at 37°C for 45 min. Released [^{14}C]O_2 was captured in a filter paper soaked in NCS II Tissue Solubilizer (GE Healthcare) suspended above the assay in a sealed 30 ml vial (VWR, Dietikon, Switzerland). Assays were stopped by addition of 100 µl ice-cold 1 M KH_2PO_4 and the filter papers were transferred to scintillation vials filled with 10 ml of IRGA-Safe Plus scintillation fluid (PerkinElmer Life Sciences). Radioactivity was measured in a Tri-Carb 2900TR scintillation counter (PerkinElmer Life Sciences).

Amino acid analysis - Purified collagens (10 µg) were hydrolyzed in 500 µl of 6 M HCl for 12 h at 105°C. Hydrolysates were dried down under nitrogen, then washed twice with 500 µl of H_2O and dried down again. Samples were resuspended in 100 µl of H_2O and derivatized using 9-fluorenylmethoxycarbonyl chloride (FMOC) following the procedure of Bank *et al*. (Bank et al. 1996). Derivatized amino acid samples were analyzed by reverse phase HPLC as described in Schegg *et al.* (Schegg et al. 2009).

Mass spectrometry - Purified collagens (2 µg) were alkylated with iodoacetamide and digested with trypsin (Shevchenko et al. 2006). Briefly, after diluting the sample in 100 mM ammonium bicarbonate, 0.1% (w/v) RapiGest (Waters, Saint-Quentin, France) and 5 mM dithiothreitol, the sample was heated for 30 min at 60°C, cooled, and alkylated in 15 mM iodoacetamide for 30 min in the dark. Proteins were digested with trypsin overnight at 37°C and acidified with trifluoroacetic acid to a final concentration of 0.5% prior to desalting using a C18 ZipTip (Millipore). Tryptic digests were subjected to reverse phase LC-MS/MS analysis using a custom packed 150 mm x 0.075 mm Magic C18- AQ, 3 µm, 200 Å, column (Bischoff GmbH, Leonberg, Germany) and an Orbitrap Velos mass spectrometer (Thermo-scientific). Peptides were separated with an 80 min gradient of

3% to 97% of a buffer containing 99.8% acetonitrile and 0.2% formic acid. Spectra were recorded in the higher energy collisional dissociation mode acquiring 10 MS/MS spectra per MS scan with a minimal signal threshold of 2000 counts. Peptides were identified and assigned using Matrix Science Mascot version 2.4.1 and verified with the Scaffold version 4 software (Proteome Software, Inc.) using the X! Tandem search engine. Variable modifications included 16 Da on methionine, proline and lysine.

Circular Dichroism - Proteins were purified by gel filtration using a Superdex 200 10/300 GL Column (GE – Healthcare). Protein fractions were concentrated in a 10 kDa Spin-X^{R} UF 500 centrifugal concentrator (Corning) in PBS, pH 7.4, and kept at 4°C at a concentration of 0.1 mg/ml prior to analysis. Human collagen type III was purchased from Sigma-Aldrich. Measurements were performed with a wavelength between 200 and 250 nm in a spectropolarimeter (J-810, Jasco) with a thermostated quartz cell of 1 mm length. Thermal stability was analyzed at 221.5 nm under heating at a rate of 0.5°C/min from 4 °C to 70°C.

Trypsin digestion of collagen - Recombinant hCOL3 (15 μg) in PBS pH 7.4 was digested with 15 ng trypsin (Roche) for 2 h at temperatures ranging from 10°C to 35°C. Digestions were stopped by addition of 2X Laemmli sample buffer and proteins were separated in 10% SDS-PAGE under reducing conditions.

Endothelial cell culture - Human umbilical vein endothelial cells (HUVEC) were cultured on 0.1% gelatin (Sigma-Aldrich), 0.1% recombinant hydroxylated hCOL3, 0.1% recombinant hCOL3 or 0.25% poly-D-lysine in ECM endothelial cell medium (ScienCell, Carlsbad, CA) at 37°C in 5% CO_2. For immunofluorescence, cells were seeded on glass cover slips at 1000 cells / cm^2, 13.3 μg coating matrix / cm^2 and cultured for 60 h. After washing twice with PBS, pH 7.4, cells were fixed with 2% paraformaldehyde for 10 min at room temperature, washed twice with 20 mM glycine in PBS and permeabilized with 1

mg/ml saponin. Cells were incubated with mouse anti-β-tubulin SAP.4G5 monoclonal antibody (Sigma-Aldrich) diluted 1:200 and labeled with rabbit anti-mouse IgG Alexa-488 (Life Technology) diluted 1:500 for 30 min. Nuclei were stained with DAPI (Biotium, Hayward, CA). Viability was assayed by methylthiazolyldiphenyl tetrazolium reduction using standard protocols (Mosmann 1983).

Sequence data - The L230 and L593 nucleotide sequences reported in this paper have the GenBank accession number NC_014649.1. The L230 and L593 protein sequences have the UniProtKB/Swiss-Prot accession numbers Q5UQC3 and Q5UP57, respectively. The nucleotide sequence of the human hCOL3 construct has the EMBL/EBI accession number HG779440.

2.1.4 Results

The genome of the giant virus mimivirus contains seven collagen-like genes and open reading frames annotated as putative lysyl and prolyl hydroxylases (Raoult et al. 2004). We have previously demonstrated that the open reading frame L230 encodes a bifunctional collagen lysyl hydroxylase and glucosyltransferase enzyme (Luther et al. 2011). To confirm the activity of the putative proyl-4-hydroxylase encoded by the open reading frame L593, we expressed a His-tagged version of the protein in *E. coli*. The 669 bp open reading frame L593 yielded a 26 kDa protein, which could be enriched on Ni^{2+} beads (Fig. 2A). The prolyl hydroxylase activity of the purified L593 protein was assayed using acceptor peptides featuring proline in sequences derived from human collagen type I, type II, adiponectin and mannose-binding lectin. The L593 protein was active as prolyl hydroxylase on the artificial peptide sequence $(GPP)_7$ and on the peptides GDRGETGPAGPPGAPGAPGAP and GLRGLQGPPGKLGPPGNPGPS derived respectively from collagen type I and mannose binding lectin, each featuring the GPP motif (Fig. 2B). By contrast, prolyl hydroxylase activity was minimal on the peptides GPMGPSGPAGARGIQGPQGPR and GIPGHPGHNGAPGRDGRDGTP derived respectively from collagen type II and adiponectin, which lack the GPP motif (Fig. 2B). The L593 prolyl 4-hydroxylase was also active on the non-collagenous peptide $(SPAP)_5$ derived from proline-rich mimivirus proteins, thus indicating that L593 was not strictly specific towards G-x-y repeats (Fig. 2B).

To assess the ability of mimivirus L230 lysyl hydroxylase and L593 prolyl 4-hydroxylase to modify collagen fragments produced in *E. coli*, we coexpressed the two mimivirus hydroxylases together with a 38 kDa fragment of human COL3A1 collagen type III. To this end, the mimivirus L230 and L593 open reading frames were expressed bicistronically under kanamycin selection and the human hCOL3 fragment on a separate plasmid under

ampicillin selection. The hCOL3 protein included 119 G-x-y repeats flanked by the N- and C-telopeptide sequences but lacking the N- and C-propeptide sequences (Fig. 1). The co-transformation of *E. coli* with the hydroxylase-containing plasmid and the human hCOL3 construct yielded expression of the three His-tagged target proteins at the expected molecular masses of 101 kDa, 37 kDa, and 26 kDa corresponding to L230, hCOL3, and L593, respectively (Fig. 3A). As a next step, the L230 and L593 hydroxylases were expressed without His-tags to enable the single enrichment of the hCOL3 protein from *E. coli* cell lysates. The expression of hCOL3 alone or together with L230 and L593 hydroxylases showed that the collagen fragment reached similar expression levels. After purification by Ni^{2+} affinity chromatography, 90 mg of collagen protein per liter of bacterial culture were routinely obtained. Of note, the coexpression with L230 and L593 hydroxylases produced a smear above the expected hCOL3 band, suggestive of a larger protein size or a decreased migration in polyacrylamide gels (Fig. 3B). This smear may also reflect heterogeneity at the level of prolyl- and lysyl hydroxylation of the recombinant protein.

The level of prolyl- and lysyl hydroxylation of the hCOL3 protein achieved by co-expression with L230 and L593 hydroxylases was determined by amino acid analysis. Native human collagen type III and the recombinant His-tagged hCOL3 protein expressed with or without L230 and L593 hydroxylases were hydrolyzed under acidic conditions, and derivatized with FMOC. The separation of FMOC-labeled amino acids by HPLC analysis showed that 54% of proline residues and 47% of lysine residues were hydroxylated in native human collagen type III (Fig. 4A). By comparison, the levels of prolyl and lysyl hydroxylation reached respectively 25% and 26% in the human hCOL3 protein coexpressed with the L593 and L230 hydroxylases (Fig. 4B). In the absence of L593 and L230 hydroxylases no prolyl and no lysyl hydroxylation were observed in the

recombinant hCOL3 protein (Fig. 4C). The efficient hydroxylation of recombinant hCOL3 indicates that substrates and co-factors required by the L593 and L230 hydroxylases are present in sufficient amounts in *E. coli* cultured in standard LB medium.

The distribution of Hyp and Hyl residues across the recombinant hCOL3 protein was investigated by mass spectrometry. The analysis of tryptic digested hCOL3 covered 92% of the sequence including 84 of 87 proline residues and all 12 lysine residues of the hCOL3 protein. The analysis of three different batches of recombinant hCOL3 protein revealed that between 66% and 83% of covered proline residues were hydroxylated (Table 1). For lysine residues, between 55% and 80% were detected as hydroxylated (Table 1). The assembly of tryptic peptides showed that hydroxylation was evenly distributed across the hCOL3 protein (Fig. 5A, Fig. S1). The mimivirus prolyl 4-hydroxylase did not appear to prefer proline residues at either the x or y position of the G-x-y motif. Several G-P-P motifs even included Hyp residues at both x and y positions. The recombinant hCOL3 protein included 12 lysine residues, of which 6 to 8 were hydroxylated (Fig. 5A). As observed for Hyp, the positions of Hyl residues within the G-x-y motif indicated that the mimivirus lysyl hydroxylase enzyme efficiently hydroxylated residues at both x and y positions. We compared the pattern of proline and lysine hydroxylation between native human collagen type III and the recombinant hCOL3 protein expressed in *E. coli*. The analysis revealed a similar distribution of hydroxylated amino acids across both polypeptide sequences (Fig. 5B). Overall, more Hyp residues were identified in the recombinant hCOL3 protein than in native collagen type III, although differences were minimal across the sequence regions surveyed. These sequences included only three lysine residues, only one of which was found to be hydroxylated in native collagen type III. By contrast, these three lysine residues were hydroxylated in the recombinant hCOL3 protein (Fig. 5B, Fig. S1). Hyl residues on recombinant hCOL3 were not further modified, for instance, by glycosylation.

We recently showed that the L230 protein is a bifunctional enzyme including both lysyl hydroxylase and Hyl glucosyltransferase domains (Luther et al. 2011). Whereas L230 efficiently converted lysine to Hyl, the enzyme failed to glycosylate the resulting Hyl residues on recombinant collagen, suggesting that the substrate UDP-Glc was not accessible in amounts sufficient to enable the L230-mediated glycosylation of recombinant collagen in *E. coli*.

The triple helical conformation and the thermal stability of the recombinant hCOL3 protein were investigated by circular dichroism. The ellipticity spectra obtained for non-hydroxylated and hydroxylated hCOL3 proteins showed the typical shape for triple helical collagen with a maximum peak around 221 nm and a negative peak below 200 nm (Fig. 6A). The changes in ellipticity at 221.5 nm during heating were monitored for non-hydroxylated and hydroxylated hCOL3 proteins to assess the thermal stability of both constructs. The triple helical conformation of the non-hydroxylated hCOL3 protein was unstable and showed an approximate 50% loss of ellipticity by 19.5°C (Fig. 6B). Since non-hydroxylated hCOL3 did not yield constant ellipticity values at low temperatures, 19.5°C however cannot be defined as true Tm value. By contrast, the hydroxylated hCOL3 protein showed a 50% loss of ellipticity by 24.3°C, indicating that hydroxylation increased the thermal stability of the construct (Fig. 6B). We also compared the degree of triple helical conformation in non-hydroxylated and hydroxylated hCOL3 by digestion with trypsin, which cleaves denatured collagen but not triple helical collagen (Bruckner and Prockop 1981). Hydroxylated hCOL3 was resistant to trypsin up to 30°C whereas non-hydroxylated hCOL3 was mostly degraded by 30°C (Fig. 6C). Both forms of hCOL3 were completely degraded by 35°C, which confirmed their low thermal stability below 37°C.

The biocompatibility of hydroxylated and non-hydroxylated recombinant hCOL3 produced in *E. coli* was assessed by using the protein as a matrix supporting the growth

of HUVEC. These cells prefer to grow on extracellular matrix proteins such as fibronectin and collagen (Smeets et al. 1992). The growth of HUVEC was compared between poly-D-lysine, bovine gelatin, recombinant non-hydroxylated hCOL3 and recombinant hydroxylated hCOL3 used as support. Cell morphology was examined by immunofluorescent staining of microtubules. When cultured on recombinant hydroxylated and non-hydroxylated hCOL3, cell viability after 60 h culture reached respectively 64% and 49% of the viability observed when cells grew on 0.1% gelatin. As expected, viability was lowest when cells were cultured on poly-D-lysine (Fig. 7A). Cell morphology assessed by staining of microtubules showed that cells were evenly spread and tightly attached to the gelatin and hydroxylated hCOL3 matrices as indicated by the large number of processes (Fig. 7B). By contrast, the amount of rounded cells was elevated when non-hydroxylated hCOL3 was applied as matrix and few cells were visible when cultured on poly-D-lysine (Fig. 7B). The compatibility of hydroxylated hCOL3 as support for the growth of HUVEC demonstrated that the recombinant protein was suitable for biological applications such as matrix-assisted cell proliferation and adhesion.

2.1.5 Discussion

The production of recombinant collagen requires post-translational modifications which are lacking in bacterial and yeast expression systems. In the present study we show that the prolyl hydroxylase L593 and the lysyl hydroxylase L230 from the giant virus mimivirus can be expressed as active enzymes in *E. coli* without any toxicity for the host cells. The coexpression of these two mimivirus hydroxylases with human collagen constructs enabled the efficient hydroxylation of proline and lysine residues across collagen requiring neither supplementation of co-factors, nor increased oxygen partial pressure.

To date, typical cost-effective and high-yield expression systems like yeasts and bacteria have not allowed the production of both prolyl and lysyl hydroxylated collagen because of the low activity of animal hydroxylases introduced in these hosts. The best results were obtained in the yeast *Pichia pastoris* expressing human prolyl 4-hydroxylase, which enabled 44% hydroxylation of proline residues on recombinant human collagen type III (Vuorela et al. 1997). However, efficient lysyl hydroxylation of collagen has not been achieved in *Pichia pastoris* so far. The degree of collagen hydroxylation is also a limiting factor for expression systems based on animal cells. Accordingly, the endogenous prolyl 4-hydroxylase and lysyl hydroxylase activities of insect cells did not yield efficient modification of recombinantly expressed collagen without co-transfection with human prolyl 4-hydroxylase subunits (Lamberg et al. 1996).

The expression of prolyl and lysyl hydroxylase enzymes derived from the giant virus mimivirus yielded degrees of hydroxylation for recombinantly expressed collagen close to those of native collagen type III. The roles of the prolyl hydroxylase L593 and lysyl hydroxylase L230 in mimivirus biology are unknown but the presence of seven collagen genes in the mimivirus genome (Raoult et al. 2004) suggests that L593 and L230 are

involved in the hydroxylation of mimivirus collagen. Indeed, we have previously shown that the mimivirus collagen-like protein L71 is hydroxylated and glycosylated *in vitro* by the L230 enzyme (17). Structurally related proteins are also found in related giant viruses, such as megavirus (Arslan et al. 2011) and moumouvirus (Yoosuf et al. 2012), which also include collagen genes in their genome. Considering their stability and activity when expressed in *E. coli*, proteins from giant viruses may represent a valuable source of enzymes for biotechnological applications, as shown here for the production of hydroxylated recombinant collagen.

The distribution of Hyp and Hyl residues on recombinant hCOL3 showed that mimivirus hydroxylases were able to hydroxylate proline and lysine in various sequence contexts. The pattern of prolyl hydroxylation showed that proline at either position x or y of the motif G-x-y could be efficiently hydroxylated. Studies performed on synthetic peptides containing Hyp at positions x or y or both demonstrated that Hyp at position y strongly increases thermal stability (Jiravanichanun et al. 2006) whereas Hyp at position x destabilizes the triple helical conformation in Gly-Hyp-y repeats (Inouye et al. 1982). The presence of Hyp at both positions x and y by contrast, further stabilized the triple helical conformation of the peptides (Berisio et al. 2004). The detection of several Hyp residues at position x on various types of collagen makes it difficult to predict the positional impact of Hyp on the thermal stability of more complex polypeptides (Bann and Bachinger 2000; Buechter et al. 2003; Song and Mechref 2013). Although early work demonstrated that Hyp occurs exclusively at position y (Fietzek and Rauterberg 1975), recent studies showed that Hyp also occurs at position x in fibrillar collagens (Song and Mechref 2013; Weis et al. 2010).

In animal cells the C-propeptide domain of collagen is important for the initiation of triple helix formation in the endoplasmic reticulum and contributes to the solubility of the

molecules along the secretory pathway (Boudko et al. 2012). The addition of trimerization domains to short collagen constructs, such as the bacteriophage T4 foldon domain at the C-terminus of a $[GPP]_{10}$ sequence, was reported to dramatically increase the thermal stability of the collagen construct (Boudko et al. 2002). Although advantageous in accelerating triple helix formation, the addition of propeptides in recombinant collagen constructs expressed in *E. coli* or *Pichia pastoris* later requires their removal for formation of fibrillar structures. This procedure is usually performed by pepsin digestion, which leaves the triple helical domain intact but also removes the short telopeptides sequences required for the registration of collagen molecules in order to form fibrils (Capaldi and Chapman 1982). Therefore, we chose to produce an hCOL3 construct devoid of propeptides but including the telopeptides necessary for fibrillogenesis. The absence of propeptides did not affect the solubility of the recombinant hCOL3 protein and simplified down-stream processing by avoiding protease digestion and removal from purified collagen.

The simplicity of this mimivirus hydroxylase expression system enables the efficient post-translational hydroxylation of proteins containing collagen domains. In addition to the family of true collagens, several collagenous proteins like adiponectin, mannose-binding lectin and the surfactant proteins A and D can now be produced as hydroxylated proteins in *E. coli*. The high yield of bacterial expression combined with a high degree of prolyl and lysyl hydroxylation provides the framework for the large-scale production of recombinant collagens for human applications, in which animal collagens represent significant risks for allergic reactions and zoonotic disease transmission.

ACKNOWLEDGMENTS

We are grateful to Dr. Peter Gehrig and Dr. Jonas Gossmann at the Functional Genomics Center Zurich for their support with mass spectrometric analyses and Marek Whitehead for endotoxin determination. This work was supported by the University of Zürich and by the Swiss National Foundation grant 310030-129633 to TH and by the Research Credit of the University of Zurich to SB.

COMPETING FINANCIAL INTERESTS

The University of Zürich has filed a patent on the application of the mimivirus hydroxylases for biotechnology purposes.

REFERENCES

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM (2011) Distant Mimivirus relative with a larger genome highlights the fundamental features of *Megaviridae*. Proc Natl Acad Sci U S A 108(42):17486-91

Bank RA, Jansen EJ, Beekman B, te Koppele JM (1996) Amino acid analysis by reverse-phase high-performance liquid chromatography: improved derivatization and detection conditions with 9-fluorenylmethyl chloroformate. Anal Biochem 240(2):167-76

Bann JG, Bachinger HP (2000) Glycosylation/Hydroxylation-induced stabilization of the collagen triple helix. 4-trans-hydroxyproline in the Xaa position can stabilize the triple helix. J Biol Chem 275(32):24466-9

Berisio R, Granata V, Vitagliano L, Zagari A (2004) Imino acids and collagen triple helix stability: characterization of collagen-like polypeptides containing Hyp-Hyp-Gly sequence repeats. J Am Chem Soc 126(37):11402-3

Boudko S, Frank S, Kammerer RA, Stetefeld J, Schulthess T, Landwehr R, Lustig A, Bachinger HP, Engel J (2002) Nucleation and propagation of the collagen triple helix in single-chain and trimerized peptides: transition from third to first order kinetics. J Mol Biol 317(3):459-70

Boudko SP, Engel J, Bachinger HP (2012) The crucial role of trimerization domains in collagen folding. Int J Biochem Cell Biol 44(1):21-32

Bruckner P, Prockop DJ (1981) Proteolytic enzymes as probes for the triple helical conformation of procollagen. Anal Biochem 110(2):360-8

Buechter DD, Paolella DN, Leslie BS, Brown MS, Mehos KA, Gruskin EA (2003) Co-translational incorporation of trans-4-hydroxyproline into recombinant proteins in bacteria. J Biol Chem 278(1):645-50

Capaldi MJ, Chapman JA (1982) The C-terminal extrahelical peptide of type I collagen and its role in fibrillogenesis in vitro. Biopolymers 21(11):2291-313

Eriksson M, Myllyharju J, Tu H, Hellman M, Kivirikko KI (1999) Evidence for 4-hydroxyproline in viral proteins. Characterization of a viral prolyl 4-hydroxylase and its peptide substrates. J Biol Chem 274(32):22131-4

Fichard A, Tillet E, Delacoux F, Garrone R, Ruggiero F (1997) Human recombinant alpha1(V) collagen chain. Homotrimeric assembly and subsequent processing. J Biol Chem 272(48):30083-7

Fietzek PP, Rauterberg J (1975) Cyanogen bromide peptides of type III collagen: first sequence analysis demonstrates homology with type I collagen. FEBS Lett 49(3):365-8

Guo J, Luo Y, Fan D, Yang B, Gao P, Ma X, Zhu C (2010) Medium optimization based on the metabolic-flux spectrum of recombinant *Escherichia coli* for high expression of human-like collagen II. Biotechnol Appl Biochem 57(2):55-62

Hyland J, Ala-Kokko L, Royce P, Steinmann B, Kivirikko KI, Myllyla R (1992) A homozygous stop codon in the lysyl hydroxylase gene in two siblings with Ehlers-Danlos syndrome type VI. Nat Genet 2(3):228-31

Inouye K, Kobayashi Y, Kyogoku Y, Kishida Y, Sakakibara S, Prockop DJ (1982) Synthesis and physical properties of (hydroxyproline-proline-glycine)10: hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix. Arch Biochem Biophys 219(1):198-203

Jiravanichanun N, Nishino N, Okuyama K (2006) Conformation of alloHyp in the Y position in the host-guest peptide with the pro-pro-gly sequence: implication of the destabilization of (Pro-alloHyp-Gly)10. Biopolymers 81(3):225-33

Lamberg A, Helaakoski T, Myllyharju J, Peltonen S, Notbohm H, Pihlajaniemi T, Kivirikko KI (1996) Characterization of human type III collagen expressed in a baculovirus system. Production of a protein with a stable triple helix requires coexpression with the two types of recombinant prolyl 4-hydroxylase subunit. J Biol Chem 271(20):11988-95

Luther KB, Hulsmeier AJ, Schegg B, Deuber SA, Raoult D, Hennet T (2011) Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. J Biol Chem 286(51):43701-9

Mosmann TR (1983) Rapid colorimetric assay for cellular growth and survival. Application to proliferation and cytotoxicity assays. J Immunol Meth 65:55-65

Myllyharju J, Kivirikko KI (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. Trends Genet 20(1):33-43.

Neubauer A, Neubauer P, Myllyharju J (2005) High-level production of human collagen prolyl 4-hydroxylase in *Escherichia coli*. Matrix Biol 24(1):59-68

Nokelainen M, Tu H, Vuorela A, Notbohm H, Kivirikko KI, Myllyharju J (2001) High-level production of human type I collagen in the yeast *Pichia pastoris*. Yeast 18(9):797-806

Pinkas DM, Ding S, Raines RT, Barron AE (2011) Tunable, post-translational hydroxylation of collagen Domains in Escherichia coli. ACS Chem Biol 6(4):320-4

Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306(5700):1344-50

Salo AM, Cox H, Farndon P, Moss C, Grindulis H, Risteli M, Robins SP, Myllyla R (2008) A connective tissue disorder caused by mutations of the lysyl hydroxylase 3 gene. Am J Hum Genet 83(4):495-503

Schegg B, Hülsmeier AJ, Rutschmann C, Maag C, Hennet T (2009) Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases. Mol Cell Biol 29(4):943-952

Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1(6):2856-60

Shoulders MD, Raines RT (2009) Collagen structure and stability. Annu Rev Biochem 78:929-58

Smeets EF, von Asmuth EJ, van der Linden CJ, Leeuwenberg JF, Buurman WA (1992) A comparison of substrates for human umbilical vein endothelial cell culture. Biotech Histochem 67(4):241-50

Song E, Mechref Y (2013) LC-MS/MS Identification of the O-Glycosylation and Hydroxylation of Amino Acid Residues of Collagen alpha-1 (II) chain from Bovine Cartilage. J Proteome Res 12(8):3599-609

Stein H, Wilensky M, Tsafrir Y, Rosenthal M, Amir R, Avraham T, Ofir K, Dgany O, Yayon A, Shoseyov O (2009) Production of bioactive, post-translationally modified, heterotrimeric, human recombinant type-I collagen in transgenic tobacco. Biomacromolecules 10(9):2640-5

Takaluoma K, Hyry M, Lantto J, Sormunen R, Bank RA, Kivirikko KI, Myllyharju J, Soininen R (2007) Tissue-specific changes in the hydroxylysine content and cross-links of collagens and alterations in fibril morphology in lysyl hydroxylase 1 knock-out mice. J Biol Chem 282(9):6588-96

Tolia NH, Joshua-Tor L (2006) Strategies for protein coexpression in *Escherichia coli*. Nat Methods 3(1):55-64

Tomita M, Ohkura N, Ito M, Kato T, Royce PM, Kitajima T (1995) Biosynthesis of recombinant human pro-alpha 1(III) chains in a baculovirus expression system: production of disulphide-bonded and non-disulphide-bonded species containing full-length triple helices. Biochem J 312 (Pt 3):847-53

van der Slot AJ, Zuurmond AM, Bardoel AF, Wijmenga C, Pruijs HE, Sillence DO, Brinckmann J, Abraham DJ, Black CM, Verzijl N, DeGroot J, Hanemaaijer R, TeKoppele JM, Huizinga TW, Bank RA (2003) Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. J Biol Chem 278(42):40967-72

Van Etten JL (2003) Unusual life style of giant chlorella viruses. Annu Rev Genet 37:153-95

Vuorela A, Myllyharju J, Nissi R, Pihlajaniemi T, Kivirikko KI (1997) Assembly of human prolyl 4-hydroxylase and type III collagen in the yeast *Pichia pastoris*: formation of a stable enzyme tetramer requires coexpression with collagen and assembly of a stable collagen requires coexpression with prolyl 4-hydroxylase. EMBO J 16(22):6702-12

Weis MA, Hudson DM, Kim L, Scott M, Wu JJ, Eyre DR (2010) Location of 3-hydroxyproline residues in collagen types I, II, III, and V/XI implies a role in fibril supramolecular assembly. J Biol Chem 285(4):2580-90

Yoosuf N, Yutin N, Colson P, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola B, Raoult D, Koonin EV (2012) Related giant viruses in distant locations and different habitats: *Acanthamoeba polyphaga* moumouvirus represents a third lineage of the *Mimiviridae* that is close to the megavirus lineage. Genome Biol Evol 4(12):1324-30

FIGURE LEGENDS

FIGURE 1. DNA and protein sequence of synthetic human hCOL3 construct. The top panel shows the codon optimized DNA sequence of the truncated human hCOL3 cDNA construct flanked by a 5' *Nco*I site and 3' *Bam*HI site (underlined). The ATG and TGA stop codon are bold and shaded. The sequence encoding the His-tag preceding the stop codon is dash-underlined. The bottom panel shows the amino acid sequence of the truncated human hCOL3 protein. The Gly-x-y collagen domain is shaded.

FIGURE 2. Bacterial expression and characterization of mimivirus L593. **A**, SDS-PAGE of mimivirus L593 expressed in *E. coli* shown as cell lysate (L) and after Ni^{2+}-affinity purification (P), either after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody. **B**, Prolyl hydroxylase activity of purified mimivirus L593 assayed on the peptide acceptors $[SPAP]_5$ (1), $[GPP]_7$ (2), GDRGETGPAGPPGAPGAPGAP from human collagen type I (3), GPMGPSGPAGARGIQGPQGPR from human collagen type II (4), GLRGLQGPPGKLGPPGNPGPS from human mannose-binding lectin (5), GIPGHPGHNGAPGRDGRDGTP from human adiponectin (6). Open bars show prolyl hydroxylase activity measured without peptide acceptor and black bars with peptide acceptors. Stars above bars indicate statistically significant activity using two-tailed paired t-test ($p<0.01$).

FIGURE 3. Coexpression of His-tagged mimivirus hydroxylases L593 and L230 with His-tagged human hCOL3 fragment. **A**, SDS-PAGE of mimivirus L593, mimivirus L230, and human hCOL3 construct expressed in *E. coli* shown as cell lysate (L) and after Ni^{2+}-

affinity purification (P), either after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody. **B**, SDS-PAGE of His-tagged human hCOL3 construct expressed alone (-) or with L593 and L230 hydroxylases (+), shown after staining with Coomassie blue R-250 or after Western blotting with anti-His_6 antibody.

FIGURE 4. Amino acid analysis of native and recombinant human hCOL3. Purified hCOL3 proteins were acid hydrolyzed and the resulting amino acids labeled with FMOC, and separated by HPLC. The positions of amino acids are indicated by the single letter code. The positions of hydroxyproline (Hyp) and hydroxylysine (Hyl) are marked by arrows. **A**, native human COL3A1; **B**, recombinant hCOL3 construct coexpressed with mimivirus L593 and L230 hydroxylases; **C**, recombinant hCOL3 construct expressed alone.

FIGURE 5. Distribution of Hyp and Hyl on recombinant human hCOL3. **A**, The occurrence of hydroxylated residues was determined by mass spectrometric analysis of tryptic digests from recombinant human hCOL3 protein. Grayed sequences represent portions of the sequences not covered in the analysis. Proline (P) and lysine (K) residues identified as hydroxylated are shaded. **B**, Comparison of Hyp and Hyl distribution on stretches of native human COL3A1 (nat) and recombinant human hCOL3 (rec) produced in *E. coli*. Proline (P) and lysine (K) residues identified as hydroxylated are shaded.

FIGURE 6. Circular dichroism of recombinant human hCOL3. **A**, Samples of purified hydroxylated (left panel) and non-hydroxylated recombinant hCOL3 protein (right panel) protein at 0.1 mg/ml were scanned between 200 and 250 nm in a

spectropolarimeter. **B**, Thermal transitions of hydroxylated (left panel) and non-hydroxylated recombinant hCOL3 protein (right panel) in PBS, pH 7.4 measured at 221.5 nm under a heating rate of 0.5°C/min from 4°C to 70°C. The T_m values were determined at the midpoints of the sigmoid curves. **C**, Trypsin digestion of hydroxylated (left panel) and non-hydroxylated recombinant hCOL3 protein (right panel); the arrowhead at the right shows the position of the hCOL3 protein.

FIGURE 7. Growth of HUVEC on recombinant human hCOL3 matrix. **A**, The viability of HUVEC seeded at 1000 cells per cm^2 was determined by reduction of methylthiazolyldiphenyl tetrazolium to formazan after 60 h incubation at 37°C on the matrices: 0.1% gelatin, 0.1% hydroxylated recombinant human hCOL3 (hCOL3-OH), 0.1% non-hydroxylated recombinant human hCOL3 (hCOL3), and 0.25% poly-D-lysine (PDL). Cell viability is expressed relatively to the values obtained for the positive control, 0.1% gelatin. Data show the mean and standard error of the mean of three experiments. All conditions tested were significantly different (p-value < 0.05) to cell viability on PDL as determined by one-way ANOVA test with Bonferroni multiple comparison. **B**, Immunofluorescence microtubule (green) and DNA (blue) staining of HUVEC grown on 0.1% gelatin, 0.1% hydroxylated recombinant human hCOL3 (hCOL3- OH), 0.1% non-hydroxylated recombinant human hCOL3 (hCOL3), and 0.25% poly-D- lysine (PDL).

Table 1. Hydroxylation efficiency of recombinant hCOL3 protein. Tryptic digests were analyzed for hydroxylation of proline (Pro) and lysine (Lys) by mass spectrometry.

	AA[a]	Coverage [%]	Pro	Hyp	Hyp/Pro [%]	Lys	Hyl	Hyl/Lys [%]
hCOL3	401		87			12		
Batch 1	346	86	80	56	70	11	6	55
Batch 2	343	86	80	53	66	11	6	55
Batch 3	291	73	63	52	83	10	8	80
Combination[b]	368	92	84	59	70	12	9	75

[a] covered amino acid length

[b] combined assembly of tryptic peptides from batches 1 to

Figure 1

```
   1  CCATGGATGT ATGATTCGTA TGATGTCAAG TCGGGTGTGG CAGTGGGTGG TCTGGCAGGC
  61  TATCCGGGTC CGGCAGGTCC GCCGGGTCCG CCGGGTCCGC CGGGTACCTC TGGTCATCCG
 121  GGTAGCCCGG GCTCTCCGGG TTATCAGGGT CCGCCGGGTG AACCGGGCCA AGCGGGTCCG
 181  AGCGGTCCGC CGGGTCCGCC GGGCGCTATT GGTCCGAGTG GCCCGGCGGG TAAAGATGGC
 241  GAATCCGGTC GTCCGGGTCG TCCGGGTGAA CGCGGCCTGC CGGGTCCGCC GGGTATTAAA
 301  GGTCCGGCAG GCATCCCGGG TTTTCCGGGT ATGAAGGGTC ACCGCGGCTT CGACGGTCGT
 361  AACGGCGAAA AAGGTGAAAC CGGTGCCCCG GGTCTGAAGG GTGAAAACGG TCTGCCGGGT
 421  GAAAATGGTG CTCCGGGTCC GATGGGTCCG CGTGGCGCGC CGGGTGAACG TGGTCGTCCG
 481  GGTCTGCCGG GTGCCGCAGG TGCCCGCGGC AACGATGGTG CACGTGGCAG TGACGGTCAG
 541  CCGGGTCCGC CGGGTCCGCC GGGGACCGCT GGTTTTCCGG GCTCCCCGGG TGCAAAAGGC
 601  GAAGTGGGTC CGGCAGGCAG TCCGGGTTCC AATGGTGCAC CGGGTCAGCG CGGCGAACCG
 661  GGTCCGCAAG GCCATGCCGG TCCGCCGGGC CCGGTTGGTC CGGCAGGCAA GAGCGGTGAT
 721  CGTGGCGAAT CTGGTCCGGC CGGTCCGGCT GGTGCCCCGG GTCCGGCCGG TAGTCGCGGC
 781  GCACCGGGTC CGCAAGGCCC GCGTGGTGAC AAAGGCGAAA CCGGTGAACG CGGCGCAGCT
 841  GGTATTAAGG GCCACCGTGG TTTCCCGGGC AATCCGGGTG CACCGGGCAG CCCGGGTCCG
 901  GCTGGCCAGC AGGGTGCCAT TGGCTCTCCG GGCCCGGCCG GTCCGCGTGG TCCGGTTGGT
 961  CCGTCAGGTC CGCCGGGTAA AGATGGCACG TCGGGTCATC CGGGTCCGAT TGGTCCGCCG
1021  GGTCCGCGTG GTAATCGCGG TGAACGTGGC TCAGAAGGTT CGCCGGGTCA CCCGGGCCAA
1081  CCTGGTCCGC CGGGTCCGCC GGGTGCTCCG GGTCCGTGCT GTGGCGGTGT TGGCGCGGCC
1141  GCAATCGCGG GCATCGGCGG CGAAAAGGCG GGCGGCTTTG CTCCGTATTA TCATCATCAC
1201  CATCACCATT GAGGATCC
```

```
   1  MYDSYDVKSG VAVGGLAGYP GPAGPPGPPG PPGTSGHPGS PGSPGYQGPP GEPGQAGPSG
  61  PPGPPGAIGP SGPAGKDGES GRPGRPGERG LPGPPGIKGP AGIPGFPGMK GHRGFDGRNG
 121  EKGETGAPGL KGENGLPGEN GAPGPMGPRG APGERGRPGL PGAAGARGND GARGSDGQPG
 181  PPGPPGTAGF PGSPGAKGEV GPAGSPGSNG APGQRGEPGP QGHAGPPGPV GPAGKSGDRG
 241  ESGPAGPAGA PGPAGSRGAP GPQGPRGDKG ETGERGAAGI KGHRGFPGNP GAPGSPGPAG
 301  QQGAIGSPGP AGPRGPVGPS GPPGKDGTSG HPGPIGPPGP RGNRGERGSE GSPGHPGQPG
 361  PPGPPGAPGP CCGGVGAAAI AGIGGEKAGG FAPYYHHHHH H
```

Figure 2

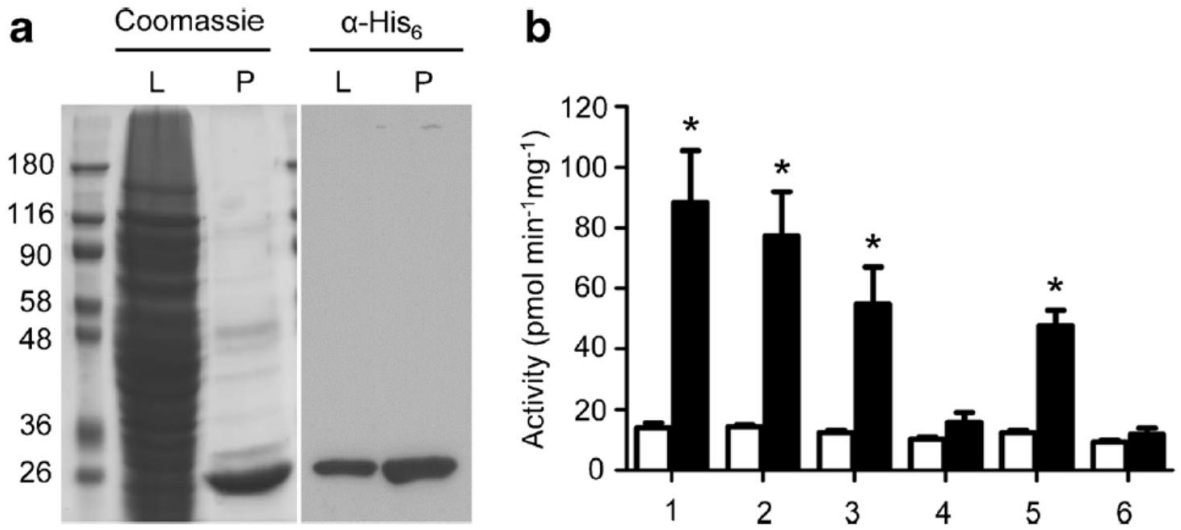


Figure 3

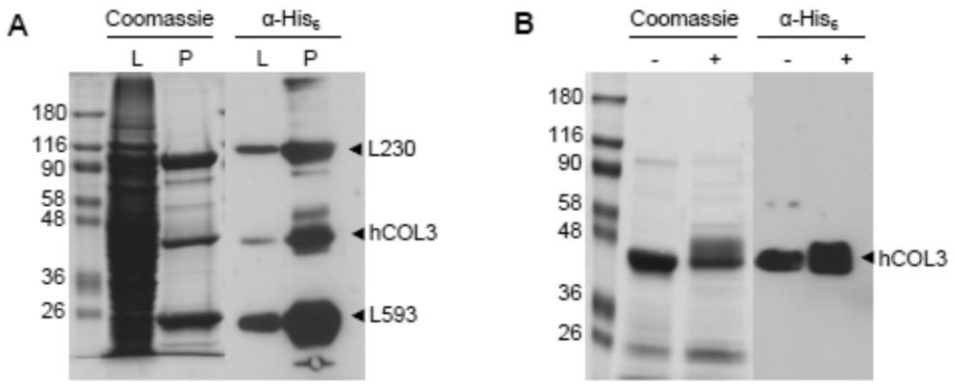


Figure 4

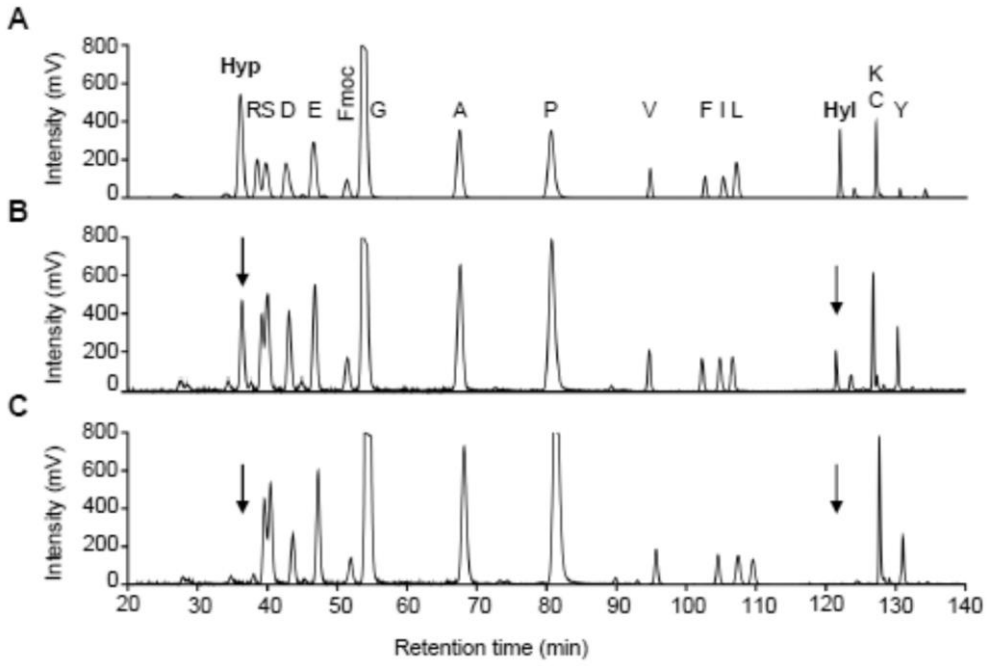


Figure 5

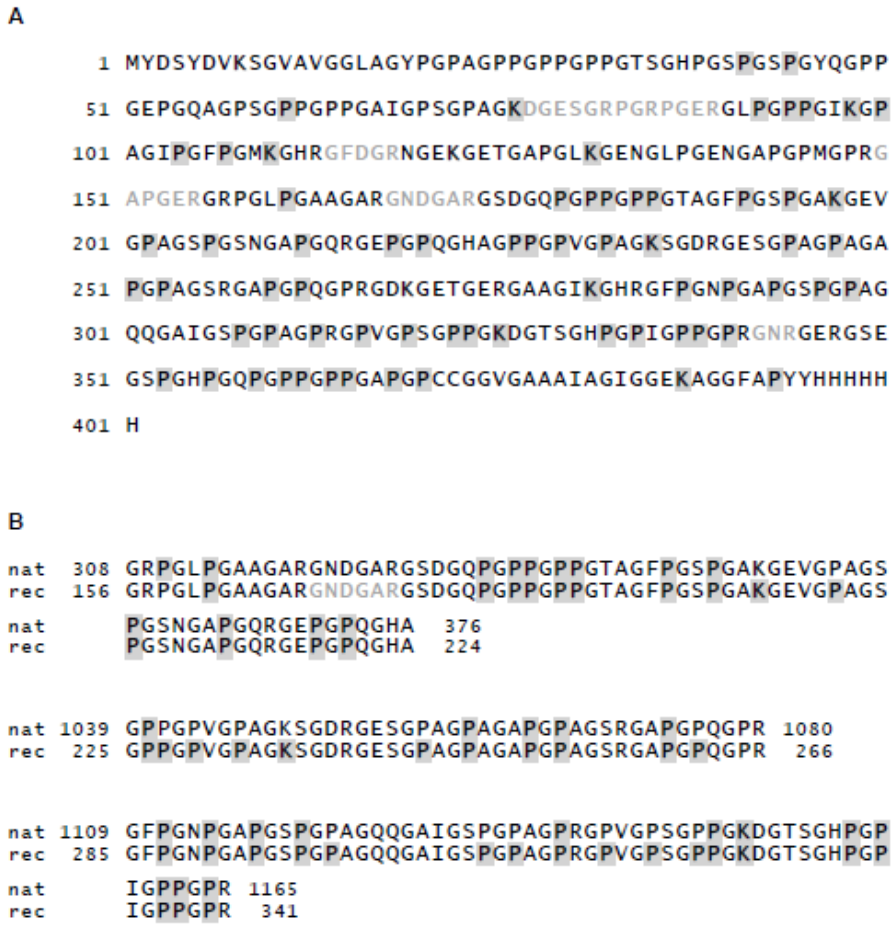

A
1 MYDSYDVKSGVAVGGLAGYPGPAGPPGPPGPPGTSGHPGSPGSPGYQGPP
51 GEPGQAGPSGPPGPPGAIGPSGPAGKDGESGRPGRPGERGLPGPPGIKGP
101 AGIPGFPGMKGHRGFDGRNGEKGETGAPGLKGENGLPGENGAPGPMGPRG
151 APGERGRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEV
201 GPAGSPGSNGAPGQRGEPGPQGHAGPPGPVGPAGKSGDRGESGPAGPAGA
251 PGPAGSRGAPGPQGPRGDKGETGERGAAGIKGHRGFPGNPGAPGSPGPAG
301 QQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGPIGPPGPRGNRGERGSE
351 GSPGHPGQPGPPGPPGAPGPCCGGVGAAAIAGIGGEKAGGFAPYYHHHHH
401 H
B
nat 308 GRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGS
rec 156 GRPGLPGAAGARGNDGARGSDGQPGPPGPPGTAGFPGSPGAKGEVGPAGS
nat PGSNGAPGQRGEPGPQGHA 376
rec PGSNGAPGQRGEPGPQGHA 224
nat 1039 GPPGPVGPAGKSGDRGESGPAGPAGAPGPAGSRGAPGPQGPR 1080
rec 225 GPPGPVGPAGKSGDRGESGPAGPAGAPGPAGSRGAPGPQGPR 266
nat 1109 GFPGNPGAPGSPGPAGQQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGP
rec 285 GFPGNPGAPGSPGPAGQQGAIGSPGPAGPRGPVGPSGPPGKDGTSGHPGP
nat IGPPGPR 1165
rec IGPPGPR 341

Figure 6

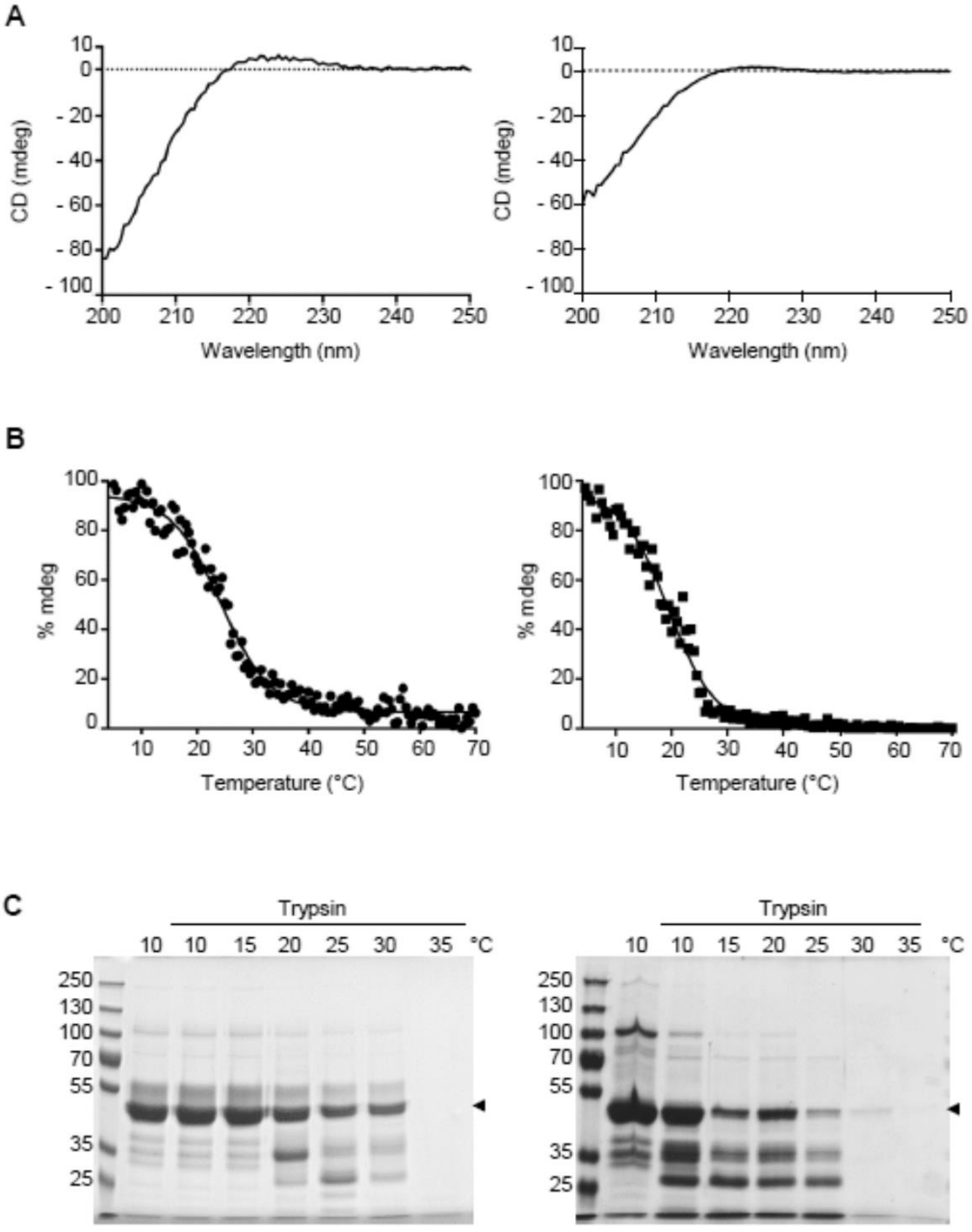
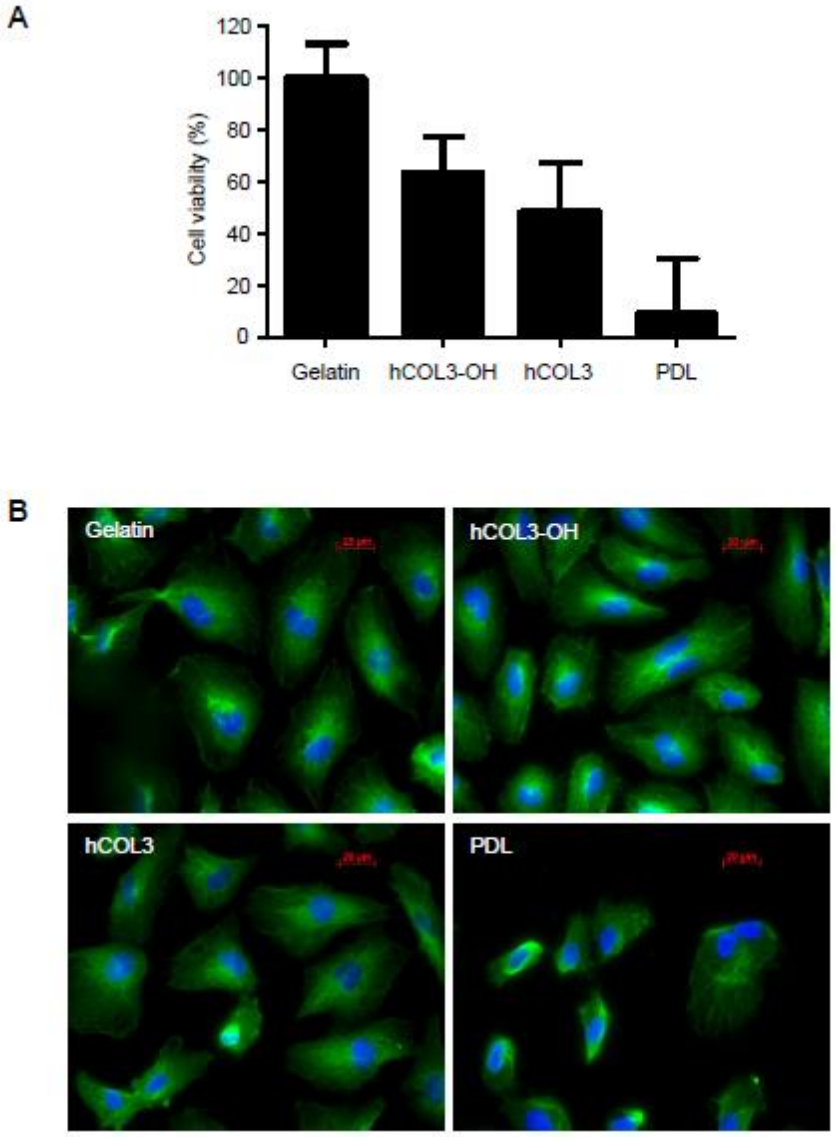

A
CD (mdeg)
Wavelength (nm)
B
% mdeg
Temperature (°C)
C
Trypsin
10 10 15 20 25 30 35 °C
250
130
100
70
55
35
25

Figure 7


A
120
100
80
60
40
20
0
Cell viability (%)
Gelatin
hCOL3-OH
hCOL3
PDL
B
Gelatin
hCOL3-OH
hCOL3
PDL

Submitted to Journal of Cell Science

2.2 COLLAGEN ACCUMULATION IN OSTEOSARCOMA CELLS LACKING GLT25D1 COLLAGEN GALACTOSYLTRANSFERASE

Stephan Baumann, Thierry Hennet

Institute of Physiology, University of Zurich, Zurich, Switzerland

Running Title: Accumulation of collagen caused by depleted glycosylation

Correspondence to: Thierry Hennet, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland, Tel. +41 44 635 50 80, thierry.hennet@uzh.ch

Number of characters in manuscript: 24’673

2.2.1 ABSTRACT

Collagen is post-translationally modified by prolyl and lysyl hydroxylation and subsequently by glycosylation of hydroxylysine. Despite the widespread occurrence of the glycan structure Glc(α1- 2)Gal linked to hydroxylysine in animals, the functional significance of collagen glycosylation remains elusive. To address the role glycosylation in collagen expression, folding and secretion, we used the CRISPR/Cas9 system to inactivate the collagen galactosyltransferase *GLT25D1* and *GLT25D2* genes in osteosarcoma cells. Loss of *GLT25D1* lead to increased expression and intracellular accumulation of collagen type I whereas loss of *GLT25D2* had no effect on collagen secretion. Inactivation of the *GLT25D1* gene resulted in a compensatory induction of *GLT25D2* expression. Loss of *GLT25D1* decreased collagen glycosylation by up to 60% but did not alter collagen folding and thermal stability. Whereas cells harboring individually inactivated *GLT25D1* and *GLT25D2* genes could be recovered and maintained in culture, cell clone with simultaneously inactive *GLT25D1* and *GLT25D2* genes could be not grown and studied, suggesting that a complete loss of collagen glycosylation impairs osteosarcoma cell proliferation and viability.

2.2.2 Introduction

Collagens, the most abundant animal proteins, are essential components of the extracellular matrix of various tissues and organs. Collagens are mainly located in connective tissue and regulate a variety of biological processes such as cell attachment, migration, proliferation, and differentiation [1].

Collagens feature specific domains composed of Gly-X-Y repeats with proline and lysine often occupying the X and Y positions. Nascent procollagen chains are modified co-translationally by prolyl 4-hydroxylation [2], prolyl 3-hydroxylation [3], lysyl hydroxylation [4], and by glycosylation of selected hydroxylysine (Hyl) residues [5]. Collagen modifications take place in the endoplasmic reticulum before completion of triple helix assembly [6]. The importance of collagen hydroxylation is underlined by various diseases linked to defective collagen modifications. For example, prolyl 3-hydroxylase 1 deficiency causes osteogenesis imperfecta with severe skeletal deformation [7], mutation in the lysyl hydroxylase 1 *PLOD1* gene causes Ehlers-Danlos syndrome type VI [8].

Mutations in the lysyl hydroxylase 2 *PLOD2* gene cause Bruck syndrome [9] and mutations in the lysyl hydroxylase 3 *PLOD3* gene cause connective tissue defects typical of collagen disorders [10]. A recently identified mutation of the prolyl 4-hydroxylase β-subunit protein disulfide isomerase causes Cole-Carpenter syndrome [11]. By contrast, the biological significance of collagen glycosylation remains elusive as no disease has been associated with the process and no model organism harboring defective collagen glycosylation has been described to date.

The *GLT25D1* and *GLT25D2* genes encode Hyl-specific galactosyltransferase enzymes, which initiate collagen glycosylation [12,13]. The gene encoding the α1-2 glucosyltransferase enzyme, which adds Glc to galactose (Gal), has not been identified yet, although the lysyl hydroxylase 3 enzyme has been claimed to also act as a collagen glucosyltransferase [14].The resulting Glc(α1-2)Gal disaccharide is strongly conserved in all animal collagens, from sponges up to mammals [14-17]. Whereas *GLT25D1* is broadly expressed across tissues, *GLT25D2* is mainly expressed in brain tissue and at low levels in skeletal muscle [18]. The GLT25D1 and GLT25D2 galactosyltransferases share identical enzymatic activities and substrate specificities, as they are able to glycosylate various types of collagen to similar

levels [18].

Despite the recent identification of the GLT25D1 and GLT25D2 galactosyltransferases, little is known about the functional role of collagen glycosylation. Glycosylation has been shown to affect the binding of the urokinase-type plasminogen activator receptor associated protein (uPARAP) to collagen type IV, thereby implying collagen glycosylation in receptor-mediated matrix remodeling [19]. Also integrins appear to be sensitive to collagen glycosylation, as decreased integrin-mediated cell adhesion was measured on galactosylated collagen peptides compared with unmodified peptides [20,21]. Glycosylation of Hyl is notably not confined to collagens. The collagen domains of multimeric proteins such as adiponectin and mannose-binding lectin also carry glycosylated Hyl, where lysyl hydroxylation and glycosylation influence protein oligomerization [22-24].

Considering the possible functional involvement of glycosylation in collagen folding and intracellular trafficking, we investigated collagen properties after inactivation of the *GLT25D1* and *GLT25D2* galactosyltransferase genes in osteosarcoma cells, which produce large amounts of fibrillar collagens, including collagen type I, V and minor amounts of collagen type III.

2.2.3 Materials and Methods

Cell lines and culture conditions – SaOS-2 cells (ATCC:HTB-85), U2OS cells (ATCC:HTB-96) and MG63 cells (ATCC:CRL-1427) were provided by Dr. Roman Muff (Sarkomzentrum Zürich, University of Zurich). Cells were grown in McCoy's 5A (Modified) medium (Thermo Fisher, Waltham, MA) containing 15% fetal bovine serum (Biochrom, Berlin, Germany) at 37°C in 5% CO_2.

Cloning and transfection of CRISPR/Cas9 vectors – The guide RNA (gRNA) sequences for *GLT25D1* exon 2 (fw: 5'-CACCGGAAGAGTTTGTACCATTCCG-3', rev: 5'- AAACCGGAATGGTACAAACTCTTC-3'), for *GLT25D2* exon 3 (fw: 5'-CACCGCCATGTGATGAAACTACGAC-3', rev: 5'-AAACGTCGTAGTTTCATCACAT-GGC-3') and for the control construct (fw: 5'- CACCGGAAGAGTTTGTACCTTTCCG-3', rev: 5'-AAACCGGAAAGGTACAAACTCTTC-3') were ligated into the BbsI sites of pSpCas9(BB) (gift from Dr. Feng Zhang, Addgene plasmid # 48139). Osteosarcoma cells were transfected using the AMAXA nucleofector kit (Lonza, Basel, Switzerland) according to the manufacturer protocol. Cells were selected for positive transfection with 0.5 µg/ml of puromycin (Santa Cruz Biotechnology Inc, Dallas, TX). Single clones were isolated and analyzed for mutations in the targeted genes using the Surveyor Nuclease assay (Integrated DNA Technologies, Coralville, IA) [25] and the primers 5'-GGAGAAGTGTCCTGTCCAGGGATAC-3', 5'- ACAGGGAACGGCTTGGGCAAAGGTC-3' for *GLT25D1* exon 2 and 5'-CCCTGATGAAATTGGACCAAAGC-3', 5'-TGCCTTTCTTAAAAAGTGGGGG-3' for *GLT25D2* exon 3. Mutations were confirmed by Sanger sequencing (Microsynth, Balgach, Switzerland).

Cloning and transfection of *GLT25D1* overexpression vector – *GLT25D1* cDNA was cloned from the pFastBacI baculovirus transfer vector from [18] in pcDNA3.1(+) (Thermo Fisher) using the NotI and XbaI restriction sites. SaOS-2 cells were transfected using AMAXA nucleofector kit (Lonza). Cells were selected 48 h after positive transfection using 2.5 µg/ml of geneticin (Thermo Fisher) for 10 days.

Collagen galactosyltransferase assay – 10^7 cells were lysed in 200 µl of Tris-buffered saline, 1% Triton X-100, pH 7.4 for 15 min on ice. Nuclei and debris were removed by spinning at 13'000 x g for 10 min at 4°C and supernatants were used as enzyme source. Bovine collagen type I was heat- denatured for

10 min at 60°C and kept on ice until use. Activity assays included 10 µl of cell lysate, 0.5 mg/ml denatured collagen acceptor, 60 µM UDP-Gal, 50'000 cpm of UDP-[^{14}C]Gal (GE Healthcare, Little Chalfont, UK), 10 mM $MnCl_2$, 20 mM NaCl, 50 mM morpholinepropanesulfonic acid (pH 7.4) and 1 mM 1,4-dithiothreitol. Reactions were incubated for 3 h at 37°C and stopped by addition of 500 µl of 5% trichloroacetic acid 5% phosphotungstic acid for 30 min on ice. Precipitated proteins were recovered on filters using a vacuum manifold, washed with 15 ml 50% ethanol and radioactivity was measured in a Tri-Carb 2900TR scintillation counter (PerkinElmer Life Sciences).

Quantitative PCR analysis – Total RNA was extracted from 5x10^6 cells using TRIzol reagent (Life Technologies, Carlsbad, CA) according to the manufacturer protocol. First strand cDNA was produced using 2 µg of total RNA and RevertAid reverse transcriptase (Life Technologies). 10 µl of 2x SsoAdvanced universal SYBR Green Supermix (Bio-Rad) were mixed with 1 µl of translated cDNA and 1 µl of 10 µM diluted primer pair (table I) in a 20 µl reaction. Specific real-time PCR primers (Table 1) for the unfolded protein response were selected from Oslowski *et al.* [26]. Primers for collagen type I *COL1A1* and collagen type V *COL5A1* mRNA quantification were designed using the primer-BLAST software [27] and encompassed exon sequences flanking at least one intron.

Collagen extraction – Osteosarcoma cells were grown to 70% confluency in DMEM containing 2% fetal calf serum, 50 µg/ml ascorbate and 50 µg/ml catalase. Ascorbate and catalase were replenished every 48 h for 8 days. Medium and cells were collected and digested at 10 µg/ml pepsin in 1 M HCl for 6 h at 22°C. Neutral pH was restored by addition of 1 M NaOH and collagens were precipitated in 50% ethanol at -20°C overnight. Collagens were resuspended in 10 ml 0.1 M acetic acid at 4°C for 48 h, then concentrated and purified using 100 kDa-cutoff Amicon ultra centrifugal filter units (Sigma- Aldrich). Collagens were diluted to 0.1 mg / ml and stored at 4°C for 72 h prior to analysis.

HPLC amino acid analysis – Purified collagens (10 µg) were hydrolyzed in 500 µl of 4 M KOH at 105°C for 20 h. Hydrolysates were neutralized using perchloric acid. Salt precipitates were removed and

supernatant dried down under nitrogen, washed twice with 500 µl of water, then dried down again. Samples were resuspended in 100 µl of water and derivatized using 9-fluorenylmethoxycarbonyl chloride following the procedure of Bank *et al.* [28]. Derivatized amino acid samples were analyzed by reverse phase HPLC [18].

Circular dichroism – Purified collagens were diluted to 0.1 mg/ml in 0.1 M acetic acid. Ellipticity was measured between 210 and 250 nm in a spectropolarimeter (J-810, Jasco) with a thermostated quartz cuvette with 1 mm length. Thermal stability was analyzed at 222 nm under heating at a rate of 0.5°C/min from 30 to 50°C.

Immunofluorescence – Cells were grown on sterile 11 mm glass coverslips for 48 h, then fixed in 4% paraformaldehyde in phosphate-buffered saline (PBS) for 15 min at room temperature, then permeabilized with 0.5% saponin for 10 min at room temperature. Cells were incubated with 5% bovine serum albumin (Sigma-Aldrich) for 1 h, then with primary antibodies diluted in PBS, 0.05% Tween, 1% bovine serum albumin for 1 h. After three wash steps in PBS, 0.05% Tween, cells were incubated with secondary antibodies for 1 h at room temperature. Cells were washed three times in PBS and nuclei were stained with DAPI (Biotium, Hayward, CA). Coverslips were mounted on ProLong Gold antifade medium (Life Technologies). Antibodies used were mouse anti-collagen type I (ab6308, Abcam, Cambridge, UK), rabbit anti-collagen type I (ab34710, Abcam), goat anti-collagen type III (ab24129, Abcam), rabbit anti-collagen type V (ab7046, Abcam), rabbit anti-GLT25D1 (ab151011, Abcam), mouse anti-PDI(RL90) (Alexis 804-012, Enzo Life Sciences, Lausen, Switzerland) and rabbit anti-giantin (ab24586, Abcam). The secondary antibodies used were goat anti-mouse 647 (ab150119, Abcam), goat anti-rabbit 488 (ab150077, Abcam) and donkey anti-goat 488 (ab150129, Abcam).

Image Acquisition and channel intensity quantification – Coverslips were imaged using a confocal laser scanning microscope type Leica TCS SP8 using a HCX PL APO CS2 oil objective lens at 63x magnification, f1.4 numerical aperture at room temperature. The fluorophores DAPI, Alexa 488 and Alexa 647 were excited sequentially at 405, 488 and 638 nm, respectively. The spectral emission detection unit was set between 415 nm and 525 nm for DAPI, between 495 nm and 590 nm for Alexa

488 and between 645 nm and 750 nm for Alexa 647. Images were recorded at a resolution of 1400 x 1400 pixels at 1000 Hz, line average = 3 and z-stacks (~25) height was set to system optimized by the software. Images were recorded using two hybrid detectors using the Leica LAS X software (Leica Microsystems AG, Heerbrugg, Switzerland). 3D images were projected to 2D using Imaris 7 (Bitplane AG, Switzerland) with the according MATLAB plugin (MathWorks, Bern, Switzerland) using MIP projection on x-y plane. Channel intensities were quantified using ImageJ 1.50B [29] as described [30].

Western Blot analysis – $5x10^6$ cells were lysed in RIPA lysis buffer for 30 min on ice, then spun for 15 min at 13000 x g at 4°C. Amounts of 20 µg of whole protein lysates were resolved on 10% SDS-PAGE. Proteins were blotted on a nitrocellulose membrane (GE Healthcare) and blocked in 5% skim milk powder in PBS for 1 h. Membranes were incubated overnight at 4°C with anti-GLT25D1 antibody at 1:250 (ab151011, Abcam), anti-collagen type I at 1:10'000 (ab138492, Abcam) or anti-β-tubulin I at 1:10'000 (Sigma-Aldrich) in Tris-buffered saline, 0.05% Tween, 1% bovine serum albumin. Secondary goat anti rabbit or goat anti mouse antibodies coupled to horseradish peroxidase at 1:10'000 (Sigma-Aldrich) were diluted in Tris-buffered saline, 0.05% Tween and incubated for 1 h at room temperature.

Pulse chase labelling of collagens – Cells were plated in 6-well plates at 2.5 x 10^5 cells per well and incubated for 24 h at 37°C. Medium was exchanged with DMEM (Sigma Aldrich) with 10% fetal bovine serum (Biochrom AG) containing 50 µg/ml ascorbate and 50 µg/ml catalase and cells were incubated overnight at 37°C. Cells were washed once in PBS, then pulsed in 1 ml of DMEM, 50 µg/ml ascorbate, 20 µCi/ml L-[^{14}C(U)]-proline (Perkin Elmer, Waltham, MA) for 4 h. Chase was initiated by changing medium to DMEM, 50 µg/ml ascorbate, 30 µg/ml L-proline. Secreted and cellular collagens were digested in the cell medium using 25 µg/ml pepsin in 1 M HCl for 2 h at 4°C. Neutral pH was restored by addition of 1 M NaOH and collagens were precipitated in 50% ethanol at -20°C overnight, then resuspended in Laemmli buffer. Collagens were separated in 9% SDS-PAGE and blotted on nitrocellulose membrane (GE Healthcare) prior to autoradiography. Band intensity was quantified using the open source software ImageJ [29] with the gel-analyzer plugin.

2.2.4 RESULTS

GLT25D1 and _GLT25D2_ inactivation in osteosarcoma cells

Collagen glycosylation is initiated by the transfer of Gal to Hyl catalyzed by GLT25D1 and GLT25D2 galactosyltransferase enzymes. To identify osteosarcoma cell lines generating high amounts of collagen galactosylation, we first analyzed *GLT25D1* and *GLT25D2* gene expression in the three collagen producing osteosarcoma cell lines SaOS-2, MG63 and U2OS [31]. As expected, GLT25D1 was the main collagen galactosyltransferase isoform expressed in osteosarcoma cells considering the restricted expression of *GLT25D2* in brain and skeletal muscle [12]. The transcript levels of *GLT25D2* represented only 1 to 4% of *GLT25D1* levels in the three cell lines investigated (Fig. 1A). As a comparison, the *PLOD3* gene encoding the lysyl hydroxylase 3 enzyme was expressed between 0.5 to 2-fold the levels of *GLT25D1* transcripts (Fig. 1A). The three collagen-modifying genes were expressed at the highest levels in SaOS-2 cells. Considering the strong expression of collagen galactosyltransferases and lysyl hydroxylase 3 genes in SaOS-2 cells, as well as the prominent production of collagen type I and highly glycosylated collagen type V in these cells [32], we primarily investigated the role of collagen glycosylation in SaOS-2 cells.

The *GLT25D1* and *GLT25D2* galactosyltransferase genes were inactivated in SaOS-2 cells using the CRISPR/Cas9 system [33]. The exons 7 to 10 of *GLT25D1* and the exons 7 to 12 of *GLT25D2* encode the GT25 domain required for galactosyltransferase activity (Fig. 1B). We therefore targeted exon 2 to disrupt *GLT25D1* and exon 3 to disrupt *GLT25D2*. To control for annealing specificity, we used a gRNA sequence nearly identical to the *GLT25D1*-targeting gRNA by including a single base mismatch (Fig. 1C). We obtained two cell clones with genomic mutations at the expected locus of *GLT25D1* and one cell clone with a mutation in *GLT25D2*. Sanger sequencing of the targeted *GLT25D1* exon 2 revealed a homozygous mutation in the first cell clone and compound heterozygous mutations in the second cell clone. Sequencing of *GLT25D2* exon 3 confirmed a homozygous point mutation at the expected genomic location (Fig. 1C). The frame-shift mutations detected in the *GLT25D1* clones yielded truncated open reading frames that lack the catalytic GT25 domain. To obtain cell clones containing both *GLT25D1*-null

and *GLT25D2*-null genes, we transfected *GLT25D1*-null cells with the CRISPR/Cas9 construct targeting the *GLT25D2* exon 3*,* which was successfully applied to disrupt *GLT25D2.* After screening 100 cell clones, we identified 14 clones carrying a single mutated *GLT25D2* allele but none that carried *GLT25D1* and *GLT25D2* genes inactivated on both alleles. SaOS-2 cells transfected with the mismatch control gRNA construct did not yield any genomic mutation in the *GLT25D1* targeted region after screening 10 cell clones. These 10 cell clones were pooled and used as control cells in subsequent experiments. The loss of GLT25D1 protein expression in SaOS-2 cells bearing inactivating mutations was confirmed by Western blotting (Fig. 1D). Changes in GLT25D2 protein levels could not be assessed because GLT25D2 remained undetectable by Western Blot analysis in the three osteosarcoma cell lines tested. The inactivation of the *GLT25D1* gene yielded a strong decrease of collagen galactosyltransferase activity down to 3-7% of the reference activity in native SaOS-2 cells and in the transfection control (Fig. 1E). Stable overexpression of a *GLT25D1* cDNA transgene *in GLT25D1*-null cells increased collagen galactosyltransferase by 6 to 8-fold. By contrast, the inactivation of *GLT25D2* in SaOS-2 cells only marginally decreased collagen galactosyltransferase activity to 92% of control values (Fig. 1E), thereby confirming GLT25D1 as the main enzyme mediating collagen galactosylation in SaOS-2. Taken together these data confirmed the successful functional inactivating role of the introduced mutations.

GLT25D1 inactivation upregulates _GLT25D2_ and _COL1A1_ mRNA levels

GLT25D1 mRNA levels were significantly reduced in *GLT25D1*-null cells indicating induction of mRNA nonsense-mediated decay [34] (Fig. 2A). *GLT25D1* mRNA levels were increased by 25% in *GLT25D2*-null cells, suggesting a possible compensatory effect by *GLT25D1* upon inactivation of *GL525D2*. Correspondingly, *GLT25D2* was upregulated 3- to 5-fold in *GLT25D1*-null cells (Fig. 2B). Interestingly, *GLT25D2* expression was also increased in *GLT25D1*-null cells overexpressing a *GLT25D1* cDNA transgene. As documented in a recent study, functional inactivation of single genes often results in compensatory overexpression by paralogous genes [35]. Whereas *GLT25D2* expression showed such a

compensatory pattern, *PLOD3* expression remained unchanged upon *GLT25D1* and *GLT25D2* inactivation (Fig. 2C), indicating that loss of collagen galactosyltransferase activity did not lead to increased lysyl hydroxylase gene expression. Surprisingly, *GLT25D1* inactivation induced a strong expression of the *COL1A1* gene encoding a collagen type I polypeptide (Fig. 2D). Normal *COL1A1* expression was restored to control levels in *GLT25D1*-null cells upon overexpression of a *GLT25D1* cDNA transgene, indicating that collagen galactosyltransferase activity was critical in regulating *COL1A1* expression. Accordingly, *COL1A1* mRNA levels were normal in *GLT25D2*-null cells (Fig. 2D). The impact of collagen galactosyltransferase activity on collagen expression was specific to *COL1A1* as *COL5A1*, encoding a collagen type V polypeptide, was insensitive to *GLT25D1* and *GLT25D2* alterations (Fig. 2E).

Collagen glycosylation in *GLT25D1*-null cells is partially restored by *GLT25D2*

To quantify the compensatory effect of GLT25D2 on collagen glycosylation in *GLT25D1*-null cells, we determined collagen post-translational modifications in endogenously produced collagen by amino acid analysis. The amount of glycosylated Hyl carrying the Glc(□1-2)Gal disaccharide was reduced in *GLT25D1*-null cells by 42 to 60% (Fig. 3A, B). Levels of free Hyl was correspondingly elevated in *GLT25D1*-null cells. Alterations of collagen folding result in over-modification of collagen because of extended exposure of collagen substrates to prolyl hydroxylases in the endoplasmic reticulum (ER) compartment [36,37]. Here, we did not detect any difference in hydroxyproline levels in *GLT25D1*- null collagen compared with collagen from control cells, suggesting a normal rate of collagen folding in *GLT25D1*-null cells.

Defects of triple helix formation because of altered post-translational modifications of collagen often impair the trafficking of collagens [7,38]. Under physiological conditions, only triple helical collagen is secreted, indicating that collagen secretion is directly proportional to triple helix formation [39]. We assessed the triple helical conformation of endogenous collagen in control and *GLT25D1*-null cells using circular dichroism. The ellipticity spectra showed the typical peak at 222 nm as well as a minimum under

200 nm in both control and *GLT25D1*-null cells (Fig. 3C). Thermal stability was determined by monitoring changes in ellipticity at 222 nm during heating collagen from 30°C to 50°C. Both collagens extracted from *GLT25D1*-null and control cell lines exhibited similar thermal properties with a normal melting temperature of 43.2°C (Fig. 3D). These results indicated that decreased collagen glycosylation did not alter collagen folding and thermal stability.

ER accumulation of collagen type I in *GLT25D1*-null cells

To confirm the increased collagen type I production in *GLT25D1*-null cells at the protein level, we first analyzed intracellular collagen amounts and distribution in control, *GLT25D1*- and *GLT25D2*-null SaOS-2 cells. In *GLT25D1*-null cells, collagen type I levels reached 150 to 190% of control values as measured by immunofluorescent channel intensity (Fig. 4A, B). Collagen type I levels normalized upon *GLT25D1* cDNA overexpression in *GLT25D1*-null cells. As noted for *COL1A1* mRNA levels, inactivation of *GLT25D2* did not alter intracellular collagen type I amounts. The increased collagen type I levels detected by immunofluorescence were further confirmed by Western Blot analysis (Fig. 4C). Collagen expression in *GLT25D2*-inactivated cells was equal to control cells and to *GLT25D1* cDNA-overexpressing *GLT25D1*-null cells (Fig. 4A), indicating a minor impact of *GLT25D2* inactivation on collagen expression.

Even though collagen type V contains 10 times more glycosylated Hyl residues than collagen type I [40], increased cellular amounts were only visible for collagen type I. Immunofluorescent staining for collagens type III and V remained unchanged in *GLT25D1*-null cells (Fig. 5A). Inactivation of *GLT25D1* may only affect collagen type I levels because this type of collagen is by far produced at highest amounts in SaOS-2 cells. To verify an impairment of collagen translocation, we colocalized collagen type I with the ER marker protein disulfide-isomerase (PDI). Elevated collagen type I levels were clearly associated with the ER compartment as shown by colocalization with PDI (Fig. 5B). Staining of SaOS-2 cells with the Golgi marker giantin showed a dilatation of Golgi stacks and co-localization with collagen type I, thus indicating the translocation of collagen type I from the ER to the Golgi (Fig. 5C).

Increased collagen type I does not induce ER stress

The intracellular accumulation of collagen type I and the occurrence of possibly improperly folded collagens could induce ER stress and thereby an unfolded protein response [41] in *GLT25D1*-null cells. To assess the possible induction of an unfolded protein response, we measured splicing of XBP1 (Fig. 6A, B), the mRNA levels of GRP78 (Fig. 6C) and ATF4 (Fig. 6D) as markers of the unfolded protein response. Tunicamycin was used as a positive control to induce the unfolded protein response [26]. Whereas tunicamycin treatment induced a robust unfolded protein response in control and *GLT25D1*-null cells, the three markers investigated remained unchanged in *GLT25D1*-null cells under conditions of collagen type I ER accumulation (Fig. 6A-D). The absence of unfolded protein response in *GLT25D1*-null cells indicated that the accumulation of collagen type I was not related to impaired collagen folding in the ER and delayed transfer to the Golgi apparatus. As the accumulation of collagen type I could be related to delayed collagen secretion, we quantified collagen production and secretion. The main types of collagen secreted by SaOS-2 cells are collagen type I and the highly glycosylated collagen type V [31]. Both control and *GLT25D1*-null cells produced and secreted these two typed of collagens normally as assessed by pulse-chase experiments (Fig. 7A,B). Collagen degradation (Fig. 7C) and secretion rates (Fig. 7D) remained normal in *GLT25D1*-null cells, thus showing that loss of *GLT25D1* did not alter collagen secretion.

2.2.5 Discussion

The inactivation of collagen galactosyltransferase *GLT25D1* and *GLT25D2* genes in osteosarcoma cells delineated the role of glycosylation in collagen expression and intracellular trafficking. Defective glycosylation had no impact on the rate of collagen secretion and triple helix thermal stability. By contrast, the loss of the main collagen galactosyltransferase isoform GLT25D1 led to ER accumulation of collagen type I in SaOS-2 cells. The intracellular accumulation of collagen type I resulted from increased collagen type I gene expression induced by low collagen galactosyltransferase activity. In addition, *GLT25D1* inactivation was compensated at the transcriptional level by induction of *GLT25D2*, which is normally hardly expressed in osteosarcoma cells.

Genetic compensation consecutive to gene inactivation is an effect that has previously been described in animals and plants such as *Arabidopsis* [35]. Accordingly, compensation by induction of paralogous genes often occurs in parallel to lethal mutations [42]. Our results on the induction of *GLT25D2* in *GLT25D1*-null cells are in agreement with such a compensatory reaction and underline the necessity to inactivate both collagen galactosyltransferase isoforms in order to obtain a complete loss of collagen glycosylation. Our failure to isolate cell clones harboring both inactive *GLT25D1* and *GLT25D2* genes supports the notion that collagen glycosylation is essential for osteosarcoma cell viability and that the induction of *GLT25D2* is required to prevent the complete loss of collagen galactosyltransferase activity when the main *GLT25D1* isoform is disrupted.

The inactivation of *GLT25D1* led to the upregulation of collagen type I expression, which is the main type of collagen produced in SaOS-2 cells. Upregulation of collagen type I in response to mutation of the post-translational machinery is also seen in fibroblasts from Bruck syndrome type 2 patients, in which mutations in *PLOD2* encoding the lysyl hydroxylase 2 isoform decrease the level of telopeptide lysyl hydroxylation and thereby the formation of telopeptide-based intermolecular crosslinks [9]. By contrast, hyper-modification of collagen, as observed in osteogenesis imperfecta caused by delayed collagen folding, does not increase collagen type I production [43,44]. The capping of Hyl by

glycosylation may act as a signal for the packaging of folded collagen into vesicular structures [45] in order to be transported to Golgi cisternae. Other classes of glycosylation, such as N-linked glycosylation, also convey signals for folded glycoproteins to depart the ER compartment [46]. Carbohydrate-binding proteins, such as ERGIC53 [47], recognize glycan chains on glycoproteins and mediate their transfer to the cis-Golgi compartment.

Collagen type I is regulated by various factors including interleukins, insulin-like growth factor-1 and transforming growth factor-β [48]. Mice deficient for β1-integrin lack feedback regulation of collagen synthesis [49]. Since glycosylation was shown to impact collagen-integrin interaction, integrins may be involved in the collagen type I upregulation seen in *GLT25D1*-null cells by modulated binding of collagen to integrin and subsequent alteration of transforming growth factor-β signaling. Another collagen interacting receptor, uPARAP, also depends on collagen glycosylation for matrix remodeling. uPARAP mediates the preferential endocytosis of glycosylated collagen IV and V [19,50]. uPARAP itself has no direct regulatory function in collagen expression since uPARAP$^{-/-}$ mice show normal collagen production. Nevertheless, uPARAP could be regulating collagen homeostasis by altered collagen uptake, thereby changing the intracellular collagen pool and making it available for intracellular feedback regulation.

Collagen glycosylation is conserved in the animal kingdom from sponges to humans [14-17], suggesting an important contribution of glycosylation to collagen properties in the extracellular matrix. In addition to the intracellular functions investigated in the present study of osteosarcoma cells, glycosylation may be involved in the organization of collagens in the extracellular space.
Glycosylation has been suggested to regulate crosslink formation in fibrillar collagens [51]. Also, the glycosylation of collagenous domains of adiponectin, serum mannose-binding lectin, and complement factor C1q has been involved in the control of subunit oligomerization [52,53]. Whereas the present study uncovered an unexpected link between collagen glycosylation and collagen type I expression in osteosarcoma cells, the inactivation of *GLT25D1* and *GLT25D2* in model organisms will be required to uncover the role of collagen glycosylation in shaping the extracellular matrix.

Acknowledgements

We thank Prof. Dr. Roman Muff for providing us with SaOS-2, MG63 and U2OS cells. This work was supported by the Research Credit of the University of Zurich to SB and by the Swiss National Foundation grant 310030_149949 to TH.

Author contributions

SB and TH designed and coordinated the study. SB performed the experiments. SB and TH wrote the manuscript.

Conflict of Interest

The authors declare no conflict of interest with the content of this article.

Literature

1. Myllyharju J, Kivirikko KI (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet* **20:** 33-43

2. Kukkola L, Hieta R, Kivirikko KI, Myllyharju J (2003) Identification and characterization of a third human, rat, and mouse collagen prolyl 4-hydroxylase isoenzyme. *J Biol Chem* **278:** 47685-47693

3. Vranka JA, Sakai LY, Bachinger HP (2004) Prolyl 3-hydroxylase 1, enzyme characterization and identification of a novel family of enzymes. *J Biol Chem* **279:** 23615-23621

4. Myllyharju J (2008) Prolyl 4-hydroxylases, key enzymes in the synthesis of collagens and regulation of the response to hypoxia, and their roles as treatment targets. *Annals of medicine* **40:** 402-417

5. Spiro RG (1967) The structure of the disaccharide unit of the renal glomerular basement membrane. *J Biol Chem* **242:** 4813-4823

6. Harwood R, Grant ME, Jackson DS (1975) Studies on the glycosylation of hydroxylysine residues during collagen biosynthesis and the subcellular localization of collagen galactosyltransferase and collagen glucosyltransferase in tendon and cartilage cells. *Biochem J* **152:** 291-302

7. Cabral WA, Chang W, Barnes AM, Weis M, Scott MA, Leikin S, Makareeva E, Kuznetsova NV, Rosenbaum KN, Tifft CJ*, et al.* (2007) Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. *Nature Genetics* **39:** 359-365

8. Brinckmann J, Acil Y, Feshchenko S, Katzer E, Brenner R, Kulozik A, Kugler S (1998) Ehlers-Danlos syndrome type VI: lysyl hydroxylase deficiency due to a novel point mutation (W612C). *Arch Dermatol Res* **290:** 181-186

9. van der Slot AJ, Zuurmond AM, Bardoel AFJ, Wijmenga C, Pruijs HEH, Sillence DO, Brinckmann J, Abraham DJ, Black CM, Verzijl N*, et al.* (2003) Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. *J Biol Chem* **278:** 40967-40972

10. Salo AM, Cox H, Farndon P, Moss C, Grindulis H, Risteli M, Robins SP, Myllyla R (2008) A connective tissue disorder caused by mutations of the lysyl hydroxylase 3 gene. *Am J Hum Gen* **83:** 495-503

11. Rauch F, Fahiminiya S, Majewski J, Carrot-Zhang J, Boudko S, Glorieux F, Mort JS, Bachinger HP, Moffatt P (2015) Cole-Carpenter Syndrome Is Caused by a Heterozygous Missense Mutation in P4HB. *Am J Hum Gen* **96:** 425-431

12. Schegg B, Hulsmeier AJ, Rutschmann C, Maag C, Hennet T (2009) Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases. *Mol Cell Biol* **29:** 943-952

13. Liefhebber JM, Punt S, Spaan WJ, van Leeuwen HC (2010) The human collagen beta(1-O)galactosyltransferase, GLT25D1, is a soluble endoplasmic reticulum localized protein. *BMC Cell Biol* **11:** 33

14. Heikkinen J, Risteli M, Wang CG, Latvala J, Rossi M, Valtavaara M, Myllyla R (2000) Lysyl hydroxylase 3 is a multifunctional protein possessing collagen glucosyltransferase activity. *J Biol Chem* **275:** 36158-36163

15. Wang C, Kovanen V, Raudasoja P, Eskelinen S, Pospiech H, Myllyla R (2009) The glycosyltransferase activities of lysyl hydroxylase 3 (LH3) in the extracellular space are important for cell growth and viability. *J Cell Mol Med* **13:** 508-521

16. Sricholpech M, Perdivara I, Nagaoka H, Yokoyama M, Tomer KB, Yamauchi M (2011) Lysyl hydroxylase 3 glucosylates galactosylhydroxylysine residues in type I collagen in osteoblast culture. *J Biol Chem* **286:** 8846-8856

17. Junqua S, Robert L, Garrone R, Pavansde.M, Vacelet J (1974) Biochemical and Morphological Studies on Collagens of Horny Sponges - Ircinia Filaments Compared to Spongines. *Connect Tissue Res* **2:** 193-203

18. Schegg B, Hulsmeier AJ, Rutschmann C, Maag C, Hennet T (2009) Core Glycosylation of Collagen Is Initiated by Two beta(1-O)Galactosyltransferases. *Mol Cell Biol* **29:** 943-952

19. Jurgensen HJ, Madsen DH, Ingvarsen S, Melander MC, Gardsvoll H, Patthy L, Engelholm LH, Behrendt N (2011) A novel functional role of collagen glycosylation: interaction with the endocytic collagen receptor uparap/ENDO180. *J Biol Chem* **286:** 32736-32748

20. Lauer-Fields JL, Malkar NB, Richet G, Drauz K, Fields GB (2003) Melanoma cell CD44 interaction with the alpha 1(IV)1263-1277 region from basement membrane collagen is modulated by ligand glycosylation. *J Biol Chem* **278:** 14321-14330

21. Stawikowski MJ, Aukszi B, Stawikowska R, Cudic M, Fields GB (2014) Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen. *J Biol Chem* **289:** 21591-21604

22. Peake PW, Hughes JT, Shen Y, Charlesworth JA (2007) Glycosylation of human adiponectin affects its conformation and stability. *J Mol Endocrinol* **39:** 45-52

23. Colley KJ, Baenziger JU (1987) Identification of the post-translational modifications of the core-specific lectin. The core-specific lectin contains hydroxyproline, hydroxylysine, and glucosylgalactosylhydroxylysine residues. *J Biol Chem* **262:** 10290-10295

24. Heise CT, Nicholls JR, Leamy CE, Wallis R (2000) Impaired secretion of rat mannose-binding protein resulting from mutations in the collagen-like domain. *Immunology* **165:** 1403-1409

25. Qiu P, Shandilya H, D'Alessio JM, O'Connor K, Durocher J, Gerard GF (2004) Mutation detection using Surveyor (TM) nuclease. *Biotechniques* **36:** 702-+

26. Oslowski CM, Urano F (2011) Measuring Er Stress and the Unfolded Protein Response Using Mammalian Tissue Culture System. *Methods Enzymol, Vol 490, Pt B* **490:** 71-92

27. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL (2012) Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13:**

28. Bank RA, Jansen EJ, Beekman B, Koppele JMT (1996) Amino acid analysis by reverse-phase high-performance liquid chromatography: Improved derivatization and detection conditions with 9-fluorenylmethyl chloroformate. *Anal Biochem* **240:** 167-176

29. (2008) ImageJ Short Reference. *Texts Comput Sci*, Book_Doi 10.1007/978-1-84628-968-2469-523

30. McCloy RA, Rogers S, Caldon CE, Lorca T, Castro A, Burgess A (2014) Partial inhibition of Cdk1 in G 2 phase overrides the SAC and decouples mitotic events. *Cell Cycle* **13:** 1400-1412

31. Pautke C, Schieker M, Tischer T, Kolk A, Neth P, Mutschler W, Milz S (2004) Characterization of osteosarcoma cell lines MG-63, Saos-2 and U-2 OS in comparison to human osteoblasts. *Anticancer Res* **24:** 3743-3748

32. McQuillan DJ, Richardson MD, Bateman JF (1995) Matrix deposition by a calcifying human osteogenic sarcoma cell line (SAOS-2). *Bone* **16:** 415-426

33. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8:** 2281-2308

34. Chang YF, Imam JS, Wilkinson MF (2007) The nonsense-mediated decay RNA surveillance pathway. *Annu Rev Biochem* **76:** 51-74

35. Rossi A, Kontarakis Z, Gerri C, Nolte H, Holper S, Kruger M, Stainier DY (2015) Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature* **524:** 230-233

36. Cabral WA, Chang W, Barnes AM, Weis M, Scott MA, Leikin S, Makareeva E, Kuznetsova NV, Rosenbaum KN, Tifft CJ*, et al.* (2007) Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta. *Nat Genet* **39:** 359-365

37. Pace JM, Kuslich CD, Willing MC, Byers PH (2001) Disruption of one intra-chain disulphide bond in the carboxyl-terminal propeptide of the proalpha1(I) chain of type I procollagen permits slow assembly and secretion of overmodified, but stable procollagen trimers and results in mild osteogenesis imperfecta. *J Med Gen* **38:** 443-449

38. Ruotsalainen H, Sipila L, Vapola M, Sormunen R, Salo AM, Uitto L, Mercer DK, Robins SP, Risteli M, Aszodi A*, et al.* (2006) Glycosylation catalyzed by lysyl hydroxylase 3 is essential for basement membranes. *J Cell Sci* **119:** 625-635

39. Yoshikawa K, Takahashi S, Imamura Y, Sado Y, Hayashi T (2001) Secretion of non-helical collagenous polypeptides of alpha1(IV) and alpha2(IV) chains upon depletion of ascorbate by cultured human cells. *J Biochem* **129:** 929-936

40. Mizuno K, Adachi E, Imamura Y, Katsumata O, Hayashi T (2001) The fibril structure of type V collagen triple helical domain. *Micron* **32:** 317-323

41. Hetz C (2012) The unfolded protein response: controlling cell fate decisions under ER stress and beyond. *Nat Rev Mol Cell Bio* **13:** 89-102

42. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* **421:** 63-66

43. Rowe DW, Shapiro JR, Poirier M, Schlesinger S (1985) Diminished Type-1 Collagen-Synthesis and Reduced Alpha-1(1) Collagen Messenger-Rna in Cultured Fibroblasts from Patients with Dominantly Inherited (Type-1) Osteogenesis Imperfecta. *J Clin Invest* **76:** 604-611

44. van Dijk FS, Cobben JM, Kariminejad A, Maugeri A, Nikkels PG, van Rijn RR, Pals G (2011) Osteogenesis Imperfecta: A Review with Clinical Examples. *Mol Syndromol* **2:** 1-20

45. Malhotra V, Erlmann P (2011) Protein export at the ER: loading big collagens into COPII carriers. *Embo J* **30:** 3475-3480

46. Helenius A, Aebi M (2001) Intracellular functions of N-linked glycans. *Science* **291:** 2364-2369

47. Nichols WC, Seligsohn U, Zivelin A, Terry VH, Hertel CE, Wheatley MA, Moussalli MJ, Hauri HP, Ciavarella N, Kaufman RJ*, et al.* (1998) Mutations in the ER-Golgi intermediate compartment protein ERGIC-53 cause combined deficiency of coagulation factors V and VIII. *Cell* **93:** 61-70

48. Rossert J, Terraz C, Dupont S (2000) Regulation of type I collagen genes expression. *Nephrol Dial Transplant* **15 Suppl 6:** 66-68

49. Gardner H, Broberg A, Pozzi A, Laato M, Heino J (1999) Absence of integrin alpha1beta1 in the mouse causes loss of feedback regulation of collagen synthesis in normal and wounded dermis. *J Cell Sci* **112 (Pt 3):** 263-272

50. Engelholm LH, List K, Netzel-Arnett S, Cukierman E, Mitola DJ, Aaronson H, Kjoller L, Larsen JK, Yamada KM, Strickland DK*, et al.* (2003) uPARAP/Endo180 is essential for cellular uptake of collagen and promotes fibroblast collagen adhesion. *J Cell Biol* **160:** 1009-1015

51. Terajima M, Perdivara I, Sricholpech M, Deguchi Y, Pleshko N, Tomer KB, Yamauchi M (2014) Glycosylation and Cross-linking in Bone Type I Collagen. *J Biol Chem* **289:** 22636-22647

52. Shinkai H, Yonemasu K (1979) Hydroxylysine-Linked Glycosides of Human Complement Subcomponent C1q and of Various Collagens. *Biochem J* **177:** 847-852

53. Hakansson K, Reid KBM (2000) Collectin structure: A review. *Prot Sci* **9:** 1607-1617

Figure Legends

Figure 1: **Characterization of *GLT25D1* and *GLT25D2* inactivation in osteosarcoma cell lines**. **A)** Real-time PCR analysis of *GLT25D1*, *GLT25D2* and *PLOD3* relative to GAPDH expression levels in SaOS-2, MG63 and U2OS cells. **B)** Representation of *GLT25D1* and *GLT25D2* gene structure. Dashes mark exons, blue dashes mark the gRNA target region, red dashes mark the glycosyltransferase coding region. **C)** Sequences of gRNAs targeting *GLT25D1* and *GLT25D2* and sequences of the targeted segment in cell clones transfected with control gRNA (C), with gRNA targeting *GLT25D1* (D1a, clone 1; D1b, clone 2), and *GLT25D2* (D2). **D)** Western blot of *GLT25D1* in *GLT25D1*-null cells with and without overexpression of *GLT25D1* cDNA *(rGLT25D1)*. **E)** Galactosyltransferase activity in lysates of control (C), *GLT25D1*-null (D1a, D1b) and *GLT25D2*-null (D2) cells with and without *GLT25D1* cDNA (rGLT25D1) overexpression.

Figure 2: **Transcription analysis of collagen and collagen modifying enzymes**. Real-time PCR analysis was performed on *GLT25D1*-null (D1a, D1b), *GLT25D2*-null (D2) and control (C) cell lines with and without *GLT25D1* (rGLT25D1) overexpression. Primers specific for **A)** *GLT25D1*, **B)** *GLT25D2*, **C)** *PLOD3*

D) *COL1A1* and **E)** *Col5A1* were used. Statistically significant differences as determined by two tailed student's *t*-test ($p<0.05$) are marked by stars.

Figure 3: **Analysis of collagen post-translational modifications and triple helical stability**. **A)** Amino acid analysis of collagens extracted from control (C) and *GLT25D1*-null cells (D1). Collagens were alkaline-hydrolysed and FMOC-labelled before separation by HPLC. Hyp: hydroxyproline, GG-Hyl: glucosylgalactosyl-hydroxylysine, Hyl: hydroxylysine. **B)** Zoom of region containing glycosylated Hyl.

C) Circular dichroism of collagens extracted from control (C) and *GLT25D1*-null (D1) cell lines. Spectra were recorded at 10°C between 210 and 250 nm in a spectropolarimeter. **D)** Thermal transition of control (C) and *GLT25D1*-null (D1) collagens in 0.1 M acetic acid. Temperature was raised from 30°C to 50°C with 0.5°C/min. *Tm* value was calculated at 50% triple helical signal.

Figure 4: **Immunofluorescent analysis**. **A)** Collagen type I and GLT25D1 staining in control (C), *GLT25D1*-null (D1a, D1b), and *GLT25D2*-null (D2) cells with and without overexpression of *GLT25D1* cDNA (+GLT25D1). White arrows mark cells expressing the transfected *GLT25D1* cDNA. Scale bar equals 10 µm. **B)** Quantification of collagen type I channel intensity based on 50 cells. Stars above bars indicate statistically significant differences based on two-tailed paired *t*-test ($p<0.05$). **C)** Western blot of collagen

type I in *GLT25D1*-null (D1a, D1b), *GLT25D2*-null (D2), control (C) and *GLT25D1* cDNA (+rGLT25D1) overexpressing cells. One representative experiment is shown (total n=3 independent experiments).

Figure 5: **Immunofluorecent analysis of collagen type III and V and colocalization of collagen type I with ER and Golgi. A)** Immunofluorescent staining of control (C) and *GLT25D1*-null (D1) cells with anti-collagen type V and anti-collagen type III antibodies (red). **B)** Colocalization of collagen type I (red) and the ER marker PDI (green). **C)** Colocalization of collagen type I (red) and the Golgi marker giantin (green). White arrows point to Golgi and collagen type I positive region.

Figure 6: **Analysis of the unfolded protein response. A-D)** Real-time PCR analysis of RNA extracted from control (C) and *GLT25D1*-null (D1) cells. Specific primers for **A)** XBP1, **B)** spliced XPB1, **C)** GRP78 and **D)** ATF4 were used with or without induction of the unfolded protein response using tunicamycin (TMC). Statistically significant differences as determined by two tailed student's *t*-test ($p<0.05$) are marked by stars (n=3 independent experiments).

Figure 7**: Analysis of collagen secretion. A-D)** Pulse chase analysis of collagens from control (C) (**A**) and *GLT25D1*-null (D1) **(B)** cells after pulse period of 4h. Collagen bands were quantified for cellular collagen (**C**) and for secreted collagens **(D)** using imageJ. One representative experiment is shown (total n=3 independent experiments).

Table I. Primer pairs used for amplification in real-time PCR reactions.

Gene	Forward primer	Reverse Primer
COL1A1	5'-GCTCGTGGAAATGATGGTGC-3'	5'-ACCCTGGGGACCTTCAGAG-3'
COL5A1	5'-CTTGGCCCAAAGAAAACCCG-3'	5'-TAGGAGAGCAGTTTCCCACG-3'
GAPDH	5'-CGCTCTCTGCTCCTCCTGTT-3'	5'-CCATGGTGTCTGAGCGATGT-3'
spliced *XBP1*	5'-TGCTGAGTCCGCAGCAGGTG-3'	5'-ATCCATGGGGAGATGTTCTGG-3'
unspliced *XBP1*	5'-CAGCACTCAGACTACGTGCA-3'	5'-TGGCCGGGTCTGCTGAGTCCG-3'
ATF4	5'-GTTCTCCAGCGACAAGGCTA-3	5'-ATCCTCCTTGCTGTTGTTGG-3'
GRP78	5'-TGTTCAACCAATTATCAGCAAACTC-3'	5'-TTCTGCTGTATCCTCTTCACCAGT-3'
GLT25D1	5'-ACTCACGCTACGAGCATGTC-3'	5'-GTGTCAGGGTTGAGGATCAG-3'
GLT25D2	5'-ACTATGGCTACCTGCCCATC-3'	5'-GGGACAACTGAGACATACTG-3'
PLOD3	5'-AGAACCTCAACGGGGCTTTA-3'	5'-CTTAGTGGGACCGTTTCCAT-3'

Figure 1

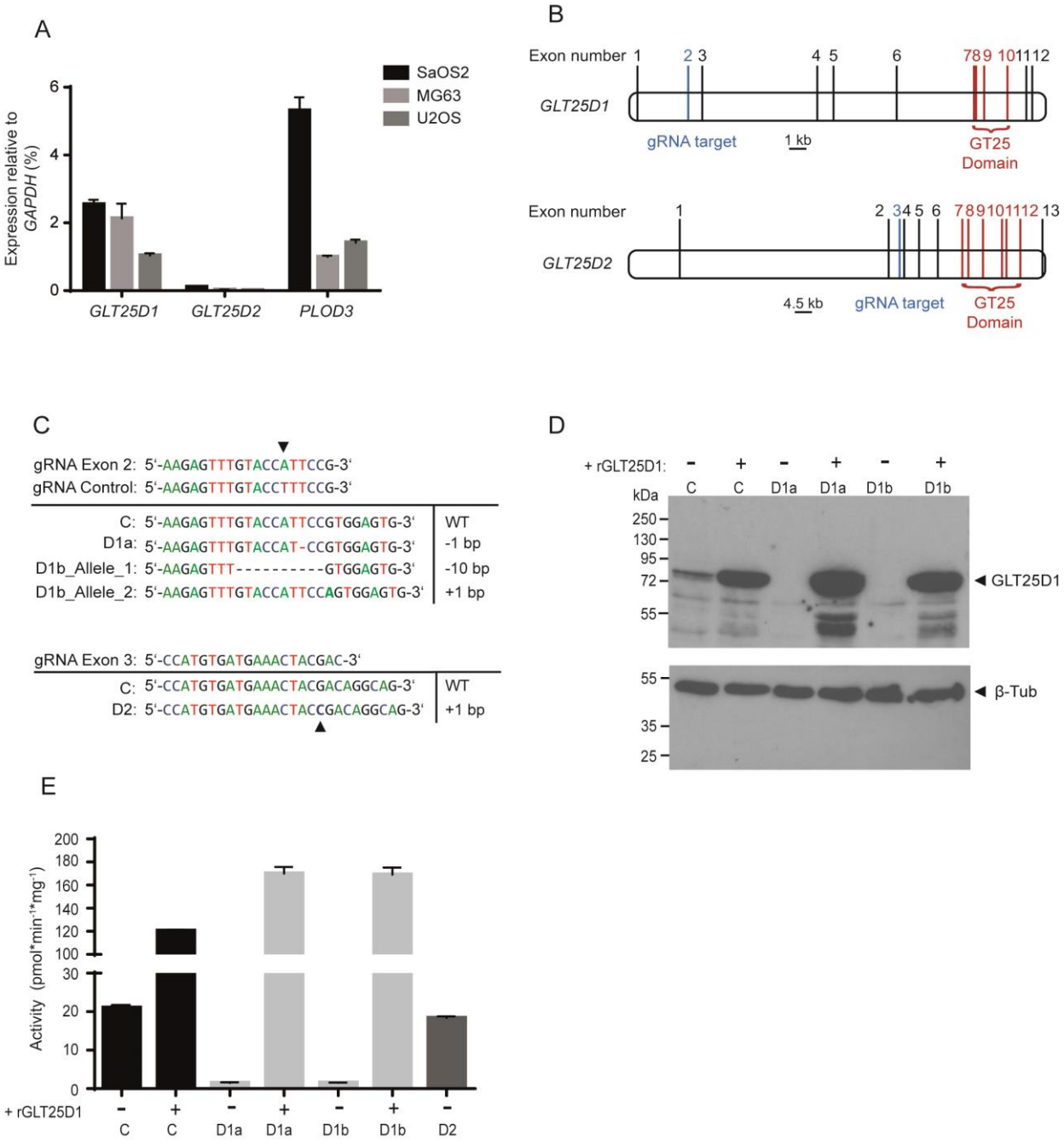

A
SaOS2
MG63
U2OS
Expression relative to GAPDH (%)
GLT25D1
GLT25D2
PLOD3
B
Exon number
GLT25D1
gRNA target
1 kb
GT25 Domain
GLT25D2
4.5 kb
C
gRNA Exon 2: 5'-AAGAGTTTGTACCATTCCG-3'
gRNA Control: 5'-AAGAGTTTGTACCTTTCCG-3'
C: 5'-AAGAGTTTGTACCATTCCGTGGAGTG-3' WT
D1a: 5'-AAGAGTTTGTACCAT-CCGTGGAGTG-3' -1 bp
D1b_Allele_1: 5'-AAGAGTTT----------GTGGAGTG-3' -10 bp
D1b_Allele_2: 5'-AAGAGTTTGTACCATTCCAGTGGAGTG-3' +1 bp
gRNA Exon 3: 5'-CCATGTGATGAAACTACGAC-3'
C: 5'-CCATGTGATGAAACTACGACAGGCAG-3' WT
D2: 5'-CCATGTGATGAAACTACCGACAGGCAG-3' +1 bp
D
+ rGLT25D1:
kDa
C C D1a D1a D1b D1b
GLT25D1
β-Tub
E
Activity (pmol*min-1*mg-1)
+ rGLT25D1
C C D1a D1a D1b D1b D2

Figure 2

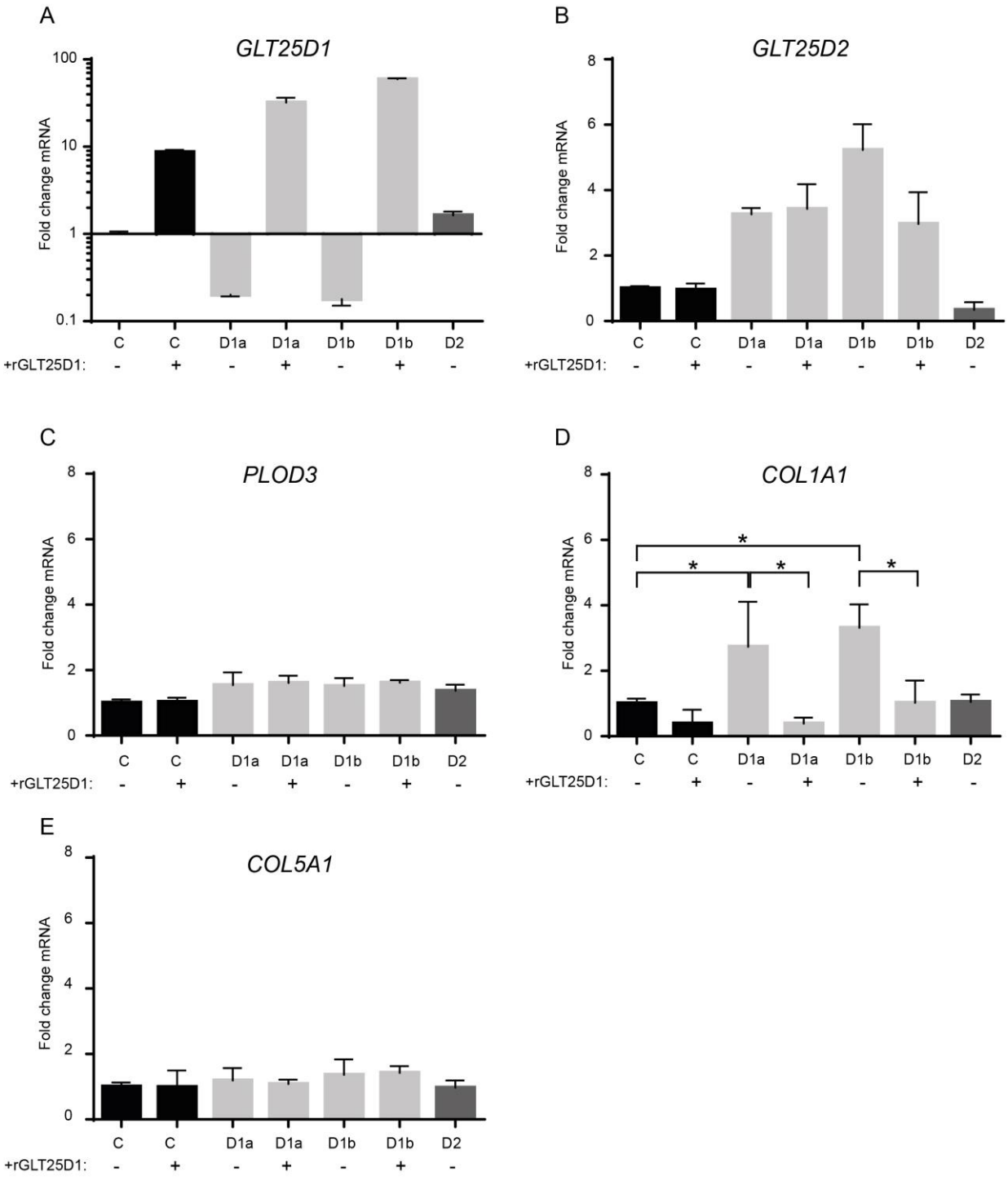

A
GLT25D1
Fold change mRNA
100
10
1
0.1
C C D1a D1a D1b D1b D2
+rGLT25D1: - + - + - + -
B
GLT25D2
Fold change mRNA
8
6
4
2
0
C C D1a D1a D1b D1b D2
+rGLT25D1: - + - + - + -
C
PLOD3
Fold change mRNA
8
6
4
2
0
C C D1a D1a D1b D1b D2
+rGLT25D1: - + - + - + -
D
COL1A1
Fold change mRNA
8
6
4
2
0
*
*
*
*
C C D1a D1a D1b D1b D2
+rGLT25D1: - + - + - + -
E
COL5A1
Fold change mRNA
8
6
4
2
0
C C D1a D1a D1b D1b D2
+rGLT25D1: - + - + - + -

Figure 3

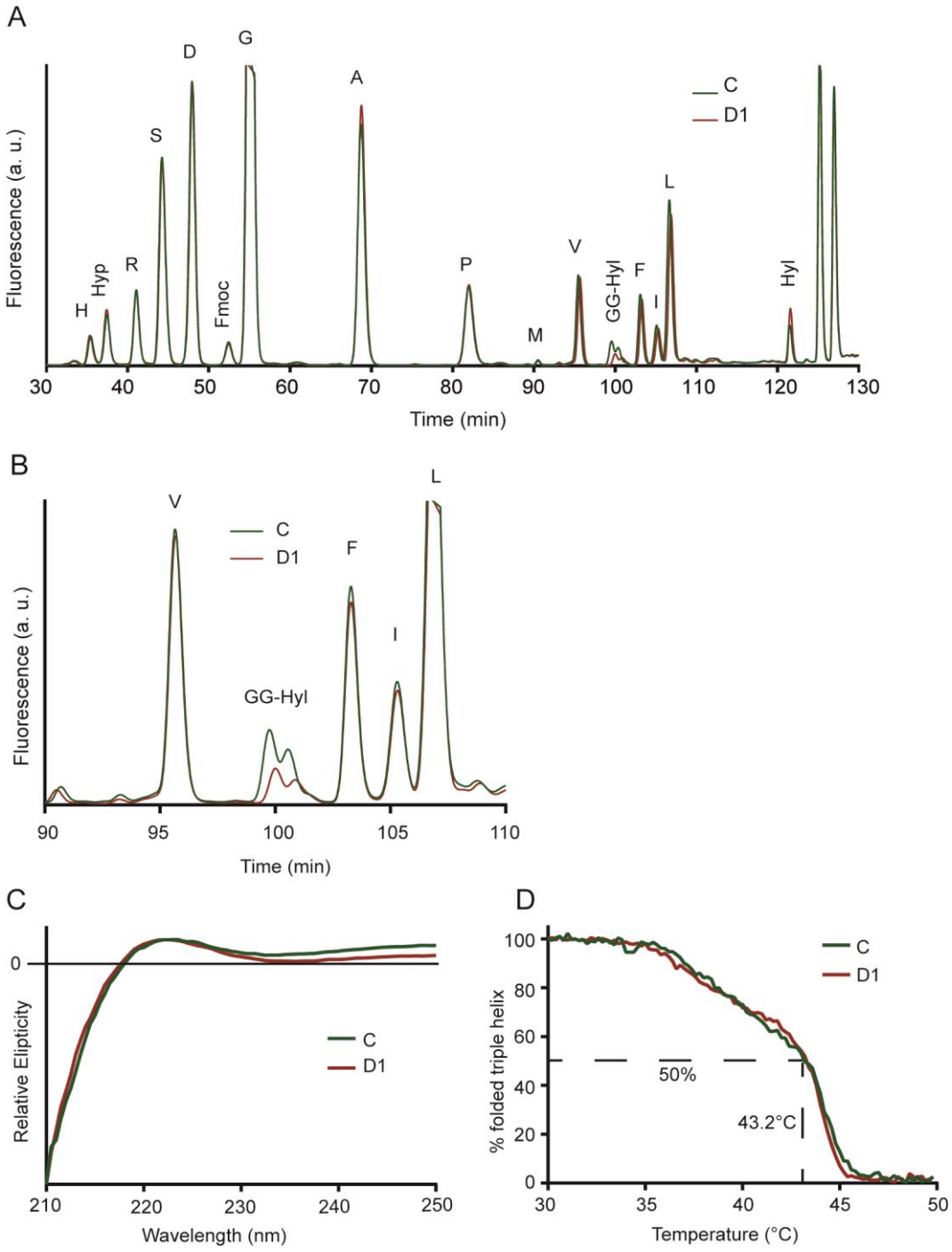

A
Fluorescence (a. u.)
H
Hyp
R
S
D
Fmoc
G
A
P
M
V
GG-Hyl
F
I
L
Hyl
C
D1
30
40
50
60
70
80
90
100
110
120
130
Time (min)
B
Fluorescence (a. u.)
V
C
D1
GG-Hyl
F
I
L
90
95
100
105
110
Time (min)
C
0
Relative Elipticity
C
D1
210
220
230
240
250
Wavelength (nm)
D
% folded triple helix
100
80
60
40
20
0
C
D1
50%
43.2°C
30
35
40
45
50
Temperature (°C)

Figure 4

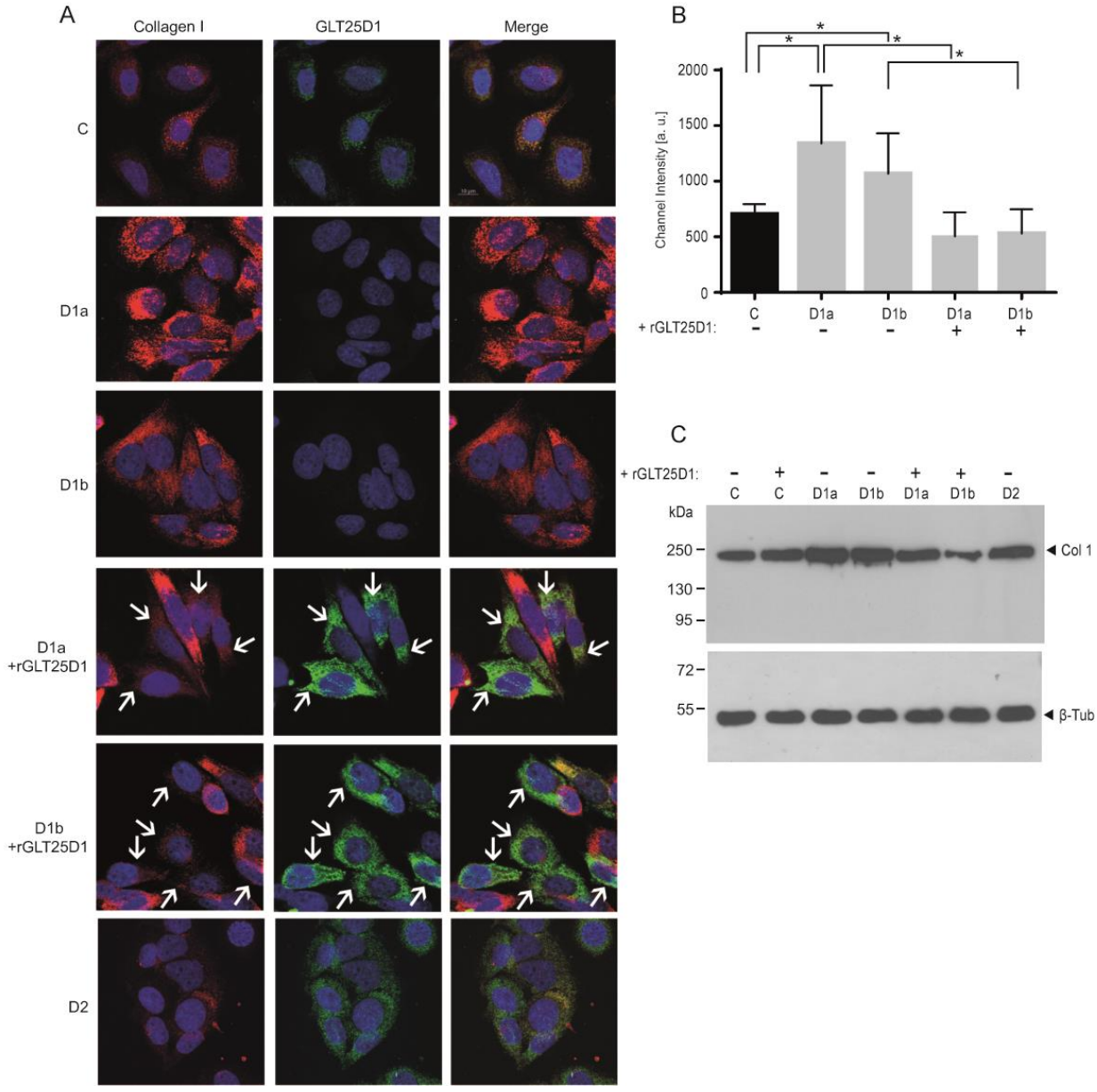

A
Collagen I
GLT25D1
Merge
C
D1a
D1b
D1a +rGLT25D1
D1b +rGLT25D1
D2
B
Channel Intensity [a. u.]
2000
1500
1000
500
0
C
D1a
D1b
D1a
D1b
+ rGLT25D1:
C
+ rGLT25D1:
C
C
D1a
D1b
D1a
D1b
D2
kDa
250
130
95
72
55
Col 1
β-Tub

Figure 5

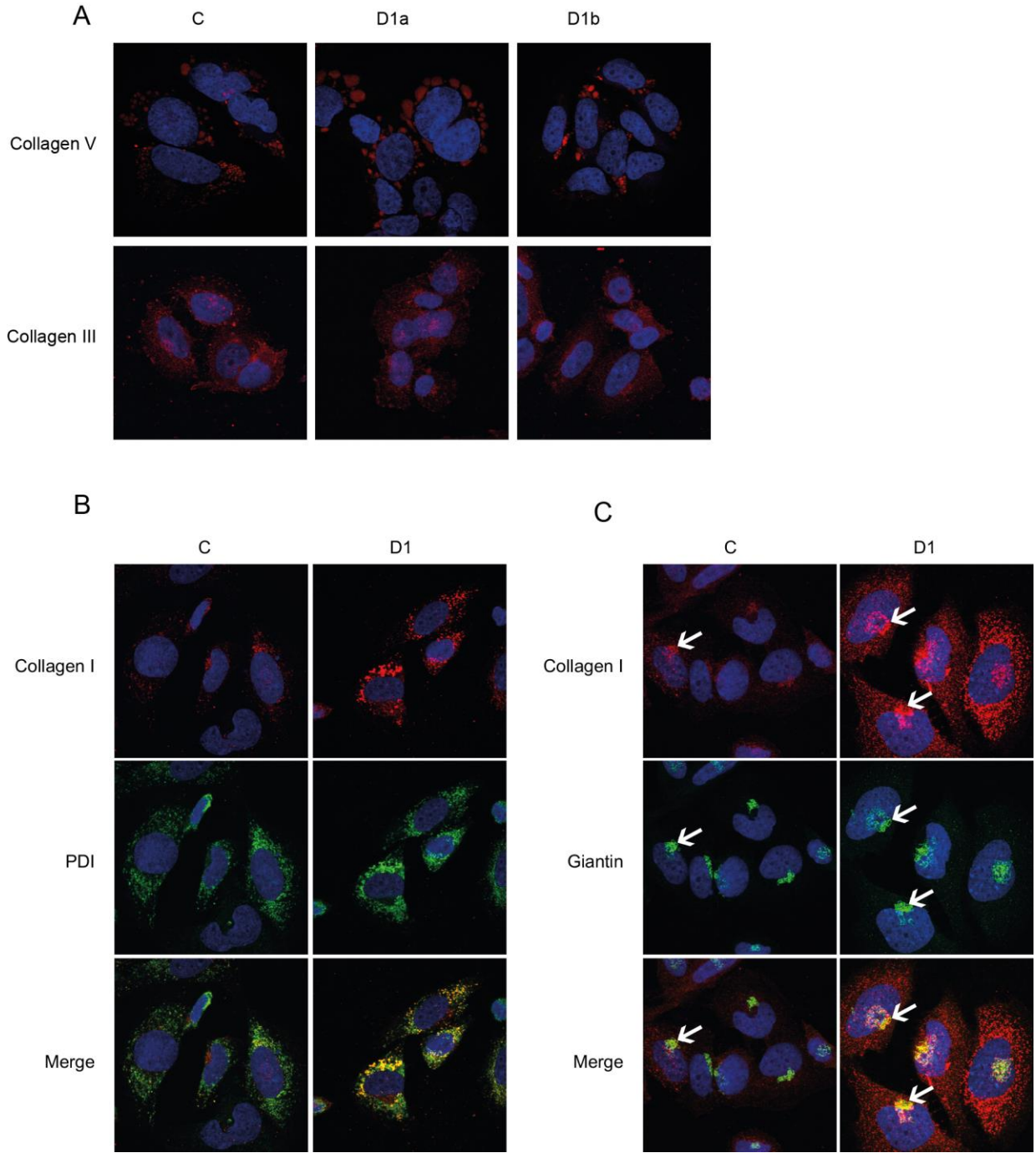

A
C
D1a
D1b
Collagen V
Collagen III
B
C
D1
Collagen I
PDI
Merge
C
C
D1
Collagen I
Giantin
Merge

Figure 6

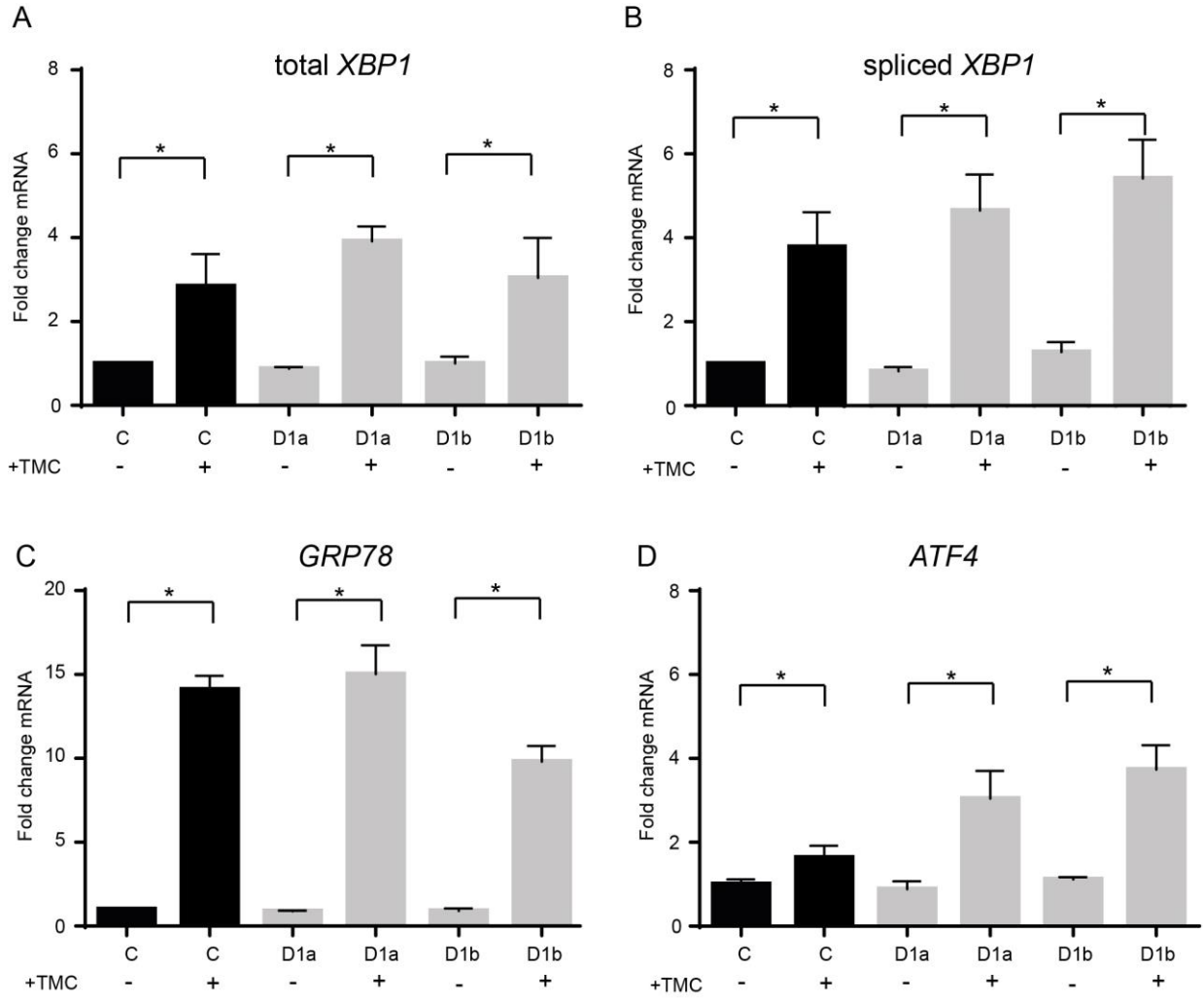
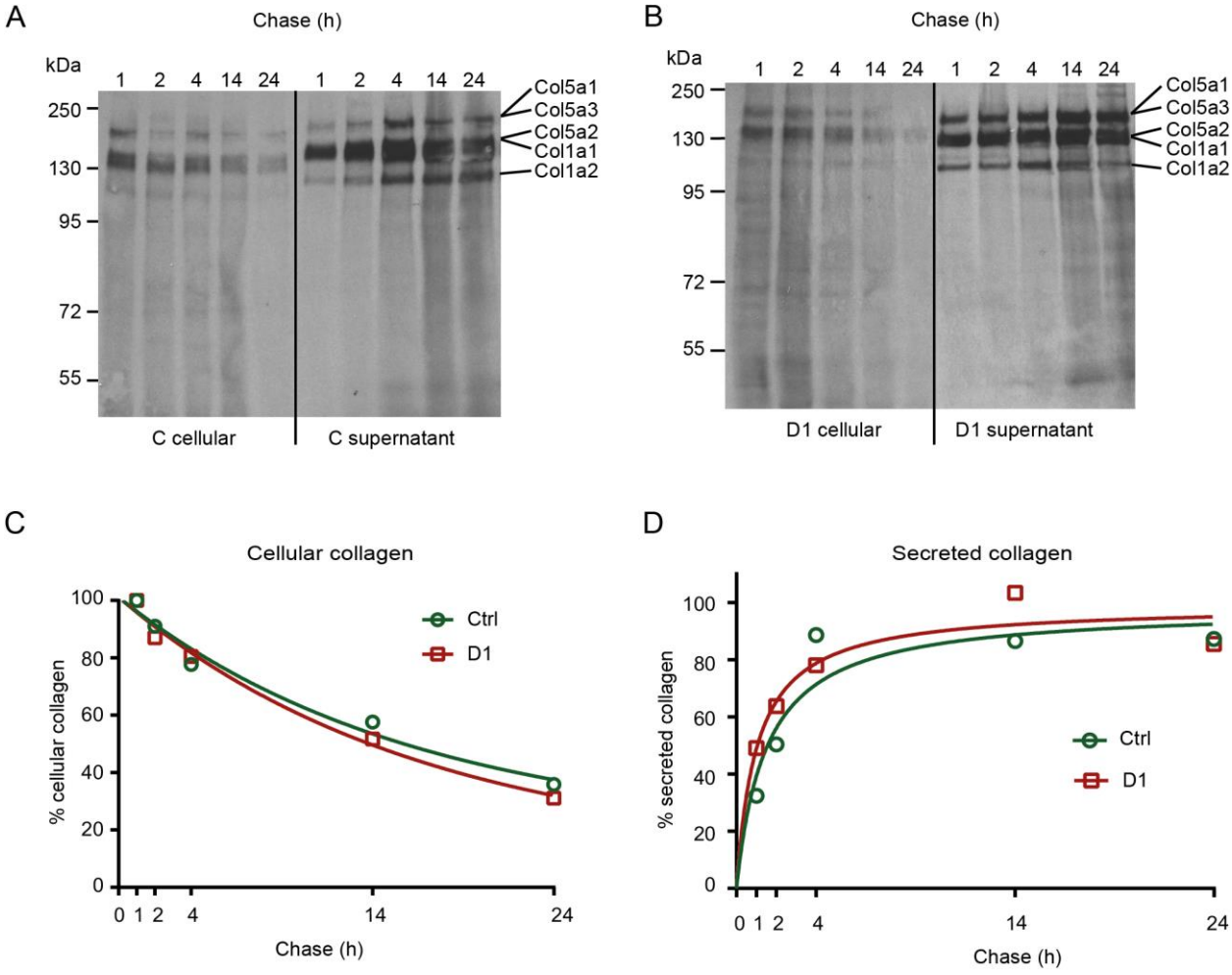

A
total XBP1
Fold change mRNA
0
2
4
6
8
*
*
*
C
C
D1a
D1a
D1b
D1b
+TMC
-
+
-
+
-
+
B
spliced XBP1
Fold change mRNA
0
2
4
6
8
*
*
*
C
C
D1a
D1a
D1b
D1b
+TMC
-
+
-
+
-
+
C
GRP78
Fold change mRNA
0
5
10
15
20
*
*
*
C
C
D1a
D1a
D1b
D1b
+TMC
-
+
-
+
-
+
D
ATF4
Fold change mRNA
0
2
4
6
8
*
*
*
C
C
D1a
D1a
D1b
D1b
+TMC
-
+
-
+
-
+

Figure 7


A
Chase (h)
kDa
1 2 4 14 24 1 2 4 14 24
250
130
95
72
55
Col5a1
Col5a3
Col5a2
Col1a1
Col1a2
C cellular
C supernatant
B
Chase (h)
kDa
1 2 4 14 24 1 2 4 14 24
250
130
95
72
55
Col5a1
Col5a3
Col5a2
Col1a1
Col1a2
D1 cellular
D1 supernatant
C
Cellular collagen
% cellular collagen
100
80
60
40
20
0
Ctrl
D1
0 1 2 4 14 24
Chase (h)
D
Secreted collagen
% secreted collagen
100
80
60
40
20
0
Ctrl
D1
0 1 2 4 14 24
Chase (h)

3. GENERAL DISCUSSION

3.1 CONCLUSIONS

In the current thesis, we describe the first functional study concerning collagen galactosyltransferases, the identification and analysis of collagens in giant viruses and the development of a bacterial expression system for posttranslationally modified human collagen using viral enzymes.

Collagen glycosylation has been poorly understood even though it was first described more than 80 years ago [176]. Glycosylation is a ubiquitous posttranslational modification but due to the elaborate methods required for analysis often neglected. Studies on the function of glycosylation require identification of the glycosyltransferase corresponding to the posttranslational modification. GLT25D1 and GLT25D2, the collagen galactosyltransferases, have only been identified in 2008 [177] and consecutive functional studies have not yet been performed. Glucosylation of galactosylated collagens was claimed to be conferred by lysyl hydroxylase 3. Lysyl hydroxylase 3 is supposedly a bifunctional enzyme accomplishing lysyl hydroxylation and collagen glucosylation. Most studies mutated either the lysyl hydroxylase domain or the glycosyltransferase domain to identify specific deterioration of either enzymatic activity. A recent finding of a connective tissue disorder led to the identification of mutations in PLOD3 either in the lysyl hydroxylase domain in one allele or in the glycosyltransferase domain in the other allele [58]. Recombinant expression of the lysyl hydroxylase 3 with mutation in the glucosyltransferase domain resulted in impaired lysyl hydroxylase activity (>50%) compared to wild type lysyl hydroxylase 3. It is hence undistinguishable whether the reported defective secretion and accumulation of collagen type IV in mutated lysyl hydroxylase 3 mice [178] arise from the glycosyltransferase activity or from the lysyl hydroxylase activity, even if there is only one point mutation inserted in the glycosyltransferase domain. Follow up studies were mostly performed on purified LH3. Thereby, glucosyltransferase activity from a possible unidentified endogenous collagen glucosyltransferase could remain undetected. This hypothesis is further fortified by the fact that organisms lacking lysyl hydroxylase 3 (i.e. sponges, chicken) still carry collagen galactosyl-glucosylation. Missing identification of the collagen galactosyltransferases combined with the supposed bi-functionality of the collagen glucosyltransferase were so far responsible for the poor characterization of collagen glycosylation.

For the first time, we described a functional study involving GLT25D1 and GLT25D2 and thereby changed collagen glycosylation in osteosarcoma cell lines. We showed a compensation mechanism in GLT25D1-null cells by GLT25D2 and upregulated collagen type I expression. Many carbohydrates are known to be crucial for receptor binding and intracellular signal transduction such as N-glycosylation in integrins [179], binding of selectins to their ligands [180] or various roles in immune responses such as the regulation of IgG function by galactosylation or sialylation [181]. Hence, it is plausible that collagen glycosylation is

involved in collagen homeostasis. Collagens are regulated by a complex network of regulatory factors. Among others, TGF-β, IFN-γ, IL-1β, IFN-γ and TNF-α are known to upregulate collagen synthesis [182]. It is therefore not possible to predict the factors being involved in glycosylation mediated collagen regulation with the available data and further experiments will be necessary.

We found an upregulation of *GLT25D2* in *GLT25D1*-null cells. Interestingly, *GLT25D2* remained upregulated even if recombinant *GLT25D1* was overexpressed in *GLT25D1*-null cells. The regulation of the galactosyltransferases is consequently not controlled on protein or mRNA levels. GLT25D1 and GLT25D2 could be regulated via an epigenetic mechanism. Epigenetic imprinting is involved in regulation of the expression of many genes [183]. An epigenetic switch can be reversible, but also cases of irreversible suppression or induction of gene expression have been described [184]. DNA methylation is one possible mechanism by which gene translation is inhibited. Promoter regions of about 40% of all human genes contain CpG islands with an atypically high frequency of CpG sites [185] which were previously described as unmethylated [186]. Recent studies, however, showed the involvement of methylation in CpG islands in promoters of e.g. tumor-suppressor genes in cancer [187] and even collagens and collagen modifying enzymes were shown to be regulated epigenetically [188]. The promoter region of *GLT25D1* and *GLT25D2* contain CpG islands (Fig. 12). Therefore, *GLT25D2* might be suppressed by methylation in wild type state (Fig. 9) and activated upon *GLT25D1* inactivation. Even though experimental evidence is missing, an irreversible epigenetic switch could explain the upregulation of *GLT25D2* in *GLT25D1*-null cells and maintenance of upregulated *GLT25D2* levels in *GLT25D1*-null cells with overexpression of recombinant GLT25D1.

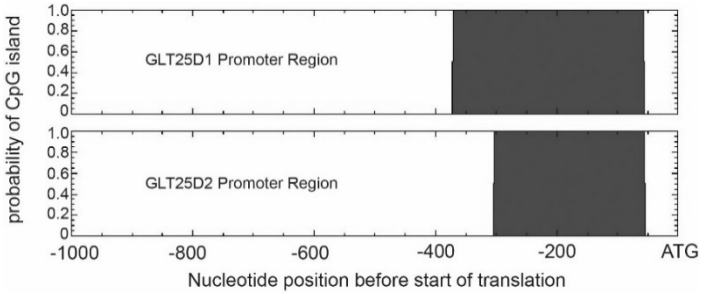


Figure 12: Identification of CpG islands in the promoter regions of GLT25D1 and GLT25D2. CpG islands are depicted in gray. The software EMBOSS Cpg-plot was used for identification.

Many collagen-modifying enzymes are essential and result in severe disorders of bones and connective tissue if mutated. Some of the diseases have been described earlier but could have been assigned only to a deficiency of specific enzymes in the last decade. Mutation in lysyl hydroxylase 1 results in the Ehlers-Danlos syndrome VI [189], mutation in lysyl hydroxylase 2 in the Bruck syndrome [190] and mutation in lysyl hydroxylase 3 in a yet unnamed collagen disorder with a combination of symptoms similar to the Ehlers-Danlos syndrome and osteogenesis imperfecta [58]. A study from 2015 further identified the first mutation in the prolyl 4-hydroxylase β-subunit resulting in the Cole-Carpenter syndrome [46]. All these

disorders combine a series of symptoms involving fragile or malformed bones, deformed and non-functional joints, but normal mental development.

In *C. elegans*, the lysyl hydroxylase 3 homologous gene let-268 leads to embryonic lethality if knocked out [191] due to missing collagen type IV secretion. In contrast, the GLT25D1 and GLT25D2 homolog D2045.9 led to a viable phenotype consisting of slower growth, aberrant locomotion and deformed mating organs [192].

Based on these findings, it is possible that a disorder of collagen glycosylation in humans exists. We identified a compensatory mechanism of galactosyltransferases by *GLT25D2* in *GLT25D1*-null cells. This compensation could partially be responsible for an understated phenotype. GLT25D2 is mainly expressed in the brain, where the stability of collagen plays a minor role. Collagen α4(V), for example, serves as an inhibitor for axonal guidance and thereby regulates the direction of axonal outgrowth in neurons [193]. A link to glycosylation has not been shown. Nevertheless, collagen glycosylation in neuronal organs could be necessary for cellular and spatial organization and could consequently be responsible for the aberrant locomotive behavior of D2045.9 knock down worms.

Collagen fibers assemble into fibrils and in supramolecular structures such as tendons, which are only formed *in vivo* and cannot be sufficiently studied *in vitro*. Glycosylation of collagen does not influence triple helix formation or their stability. Nevertheless, glycosylation could be involved in quaternary fibril formation, and its physiological role could become evident by studies on *GLT25D1* in complex organisms as zebrafish or mice. There are currently 28 different types of collagen identified, of which only a fraction has been characterized. Further studies are essential to assess the diverse possible roles of glycosylation in the various collagens.

We further identified 142 collagen-like proteins in the genomes of 60 analyzed nucleocytoplasmic large DNA viruses. Collagen-like genes from the *Megaviridae* and *Mimiviridae* families originated partially via gene duplication and share a common ancestor. In contrast, collagens from pithovirus and *Pandoraviridae* evolved via different routes and are not related. Many of the vertebrate and invertebrate collagens consist of a conserved, defined length of the major collagen domain of 1011 – 1020 amino acids containing 337 – 340 GLY-X-Y repeats [194]. The giant viruses encode longer and shorter collagen domains, but none of 337 – 340 GLY-X-Y repeat length. This finding supports the hypothesis that viral and metazoan collagens have developed co-evolutionary and exhibit different ancestors. Fibrillar collagens from early multicellular organisms as from sponge to humans share a common structure at the genomic level. The collagenous domain is encoded by exons of 54 or 45 bp length or of multiples of these modules [195, 196]. The modular architecture of fibrillar collagen indicates the existence of an ancient prototypic collagen, which was later expanded via duplications of the 54 bp exon and via unequal crossovers resulting in 45 bp exons [197]. All exons start with an intact codon for glycine and end with an intact

codon for any other amino acid. Even though giant viruses contain genomes devoid of intervening sequences, a similar evolution of the viral collagens to the evolution of metazoan collagens is possible. This hypothesis would explain the occurrence of copies of the strongly conserved interdomains. However, more DNA sequence data must be generated to identify an organism or virus containing the ancestral collagen. *Megaviridae* and *Mimiviridae* collagens share a high content of oppositely charged residues at the X and Y position of Gly-X-Y repeats. In humans, charged amino acids are important for the stabilization of fibrillar collagens with interrupted collagen domains. The length and amino acid composition of the viral collagen domains however do not confirm an evolutionary relationship between viral and human collagens. Currently only very little information is available about collagens in giant viruses. Therefore, the functions of viral collagens can be only hypothesized. In bacteria and in virophages, collagen-like proteins occur in fibers attached to the cell wall. A similar function could be hypothesized for mimivirus collagens. The mimivirus consists of a protein- and DNA-containing core, which is encompassed by two lipid membranes and a virus capsid. Several fibers are attached to the capsid [198]. The fibers are collagenase resistant indicating a composition of non-collagenous proteins [198]. However, proline is often essential in the recognition motif of collagenases to cleave collagen [199]. Considering the low frequency of proline in the viral proteins together with the high moiety of charged amino acids, it is doubtful whether the conventional collagenases can digest mimivirus collagens. The ORF L71 of mimivirus was shown to encode a collagen-like protein occurring on the surface of the virus [152]. It is currently unknown, whether L71 is part of the viral membrane or of the fibers.

Giant viruses not only contain collagen-like proteins but also collagen modifying enzymes. We used a prolyl 4-hydroxylase and a bifunctional lysyl hydroxylase and glucosylatransferase from mimivirus to express recombinant, posttranslationally modified human collagen type III in a bacterial expression system. The collagens folded triple helically and the stability of triple helical collagen could be increased by hydroxylation.

Collagen is used in reconstructive medicine and in cosmetics. To date, collagen of bovine or porcine origin is used due to a high biocompatibility. The largest segment of the market of collagens consisted of dermal fillers. The American society of Aesthetic Plastic Surgery Practice Survey from 2014 states about 22’049 collagen based applications yearly in the USA. In 2000, in contrast, there were almost 600’000 applications registered [200]. The main reason of this decay lays in the insufficient biosafety of animal collagens and in the development of safe alternatives. Bovine collagen type I shares 95% sequence identity to human collagen type I and results in allergic reactions in 1 – 3% of the applications. In contrast, hyaluronic acid is a polysaccharide with the structure $(\text{-4GlcUA}\beta\text{1-3GlcNAc}\beta\text{1-})_n$ that is conserved among various species [201]. It is therefore non-immunogenic and the most used alternative to collagen as dermal filler today with almost 2’000’000 yearly applications in the USA. Hyaluronic acid occurs in rooster combs and is mainly purified from this animal source. Even after purification, avian protein impurities can

cause allergic reactions in patients treated with hyaluronic acid [202]. With the development of a recombinant expression system for hyaluronic acid in streptococcus, large scale production became possible [203]. These non-animal stabilized hyaluronic acid products exhibit a low risk for hypersensitivity reactions (<0.6%) and are therefore commonly used today. Due to the polar composition of hyaluronic acid, it can bind approximately 6 liters of water per gram [204]. Contrary to collagen fillers, which are the main structural component in collagen-based dermal fillers and provide tensile strength, hyaluronic acid applications are stabilized by water to replenish degraded connective tissue and provide pressure resistance. Hyaluronic acid can be chemically crosslinked in order to improve half-life from three to six months.

With the advantages of less immunogenicity and enhanced durability compared to collagen fillers, hyaluronic acid is the preferred agent for dermal injections today. Nevertheless, there are novel formulations of crosslinked collagens which can result in a half-life of up to 12 months [205]. The formulation uses ribose to crosslink collagen and thereby making it less available for collagenases and cellular degradation. Since it is derived from pork, hypersensitivity reactions can still occur but they are significantly reduced compared to native porcine collagen.

The market for dermal fillers is heavily growing and novel products based on recombinant collagens could colonize a niche by implementing natural advantages of collagen. Endogenous collagen in skin exhibits a half-life of approximately 15 years [206]. In contrast, endogenous hyaluronic acid in synovial fluids exhibits a half-life of only 16 hours [207]. Even if these results do not represent the stability of dermally injected polymers, manufacturing of collagen-based devices should be possible with superior durability compared to hyaluronic acid. One approach could be implemented by the stabilization of the triple helix by prolyl 4-hydroxylation. Fully hydroxylated collagen type I exhibits a melting temperature 16°C higher compared to unhydroxylated collagen [20]. The collagen triple helices might further be cross-linked via the lysyl oxidase. These modifications should enable the production of durable collagen based applications. Moreover, collagens are hemostatic [208] resulting in less bleeding and reduced swelling after injection compared to hyaluronic acid. In addition collagen increases cell attachment, cell proliferation and migration [209]. Therefore, collagen would be the preferred substrate for intradermal applications, if its longevity could be increased.

Our study provides the proof of principle that production of posttranslationally modified recombinant human collagens in bacteria is possible. For a successful application in mice or human, several hurdles still have to be taken. The collagen expressed in our system is a truncated 38 kDa version of type III collagen. The collagen domain of full-length collagen type III however is 98 kDa and was expressed only in small amounts in *E. coli* (data not shown). By truncation, new immunogenic epitopes could be generated and biosafety compromised. The bacterially expressed collagens were only analyzed on the

level of triple-helical assembly, fibril and fiber formation have not been studied. We further used a viral prolyl 4-hydroxylase with different specificity compared to the human prolyl 4-hydroxylase. Hydroxylation of prolines in the X position of the GLY-X-Y repeats was shown to destabilize the triple helix. For maximal stability of the collagen triple helix, either a different prolyl 4-hydroxylase must be found, or the mimivirus prolyl 4-hydroxylase must be genetically modified in order to increase its specificity.

To sum up, recombinant human collagen from bacterial expression harbors an enormous potential in cosmetic and reconstructive medicine due to high expression levels and presumably low antigenicity. Preliminary hurdles such as high costs of development inhibited successful commercial applications so far. Therefore, further research on intradermal longevity and immunogenicity are indispensable.

3.2 FUTURE DIRECTIONS

The findings derived from the present study on collagen glycosylation lead to several new questions. Collagen glycosylation is still poorly understood. Detailed characterization of the regulation of collagen modifying enzymes could reveal further involvements of collagen glycosylation in biological processes such as cell division, migration or spatial organization. We presented a link between collagen glycosylation and collagen expression. Further experiments could aim on a mechanistic explanation for that phenomenon by identifying proteins interacting with glycosylated collagen but not with non-glycosylated collagen. Our experiments revealed new insights on collagen glycosylation on cellular and molecular levels. Animal models, however, could reveal novel functions of collagen glycosylation in a broader context during embryogenesis and, if viable, during early stages of life. Collagen glycosylation could also be involved in diseases of connective tissue or bone in humans. By DNA sequencing of available samples from patients suffering from unidentified disorders, mutations in *GLT25D1* or *GLT25D2* could be identified and further characterized.

We identified collagen-like proteins from giant viruses with features not observed in other collagens. The mimi- and megavirus derived collagen-like proteins have oppositely charged amino acid residues in about 60% of all G-X-Y repeats. While human collagens use a similar strategy to stabilize short, interrupted collagen domains, the viruses could also make use of the ionic interactions for triple helix stabilization. To prove this hypothesis, collagen-like proteins can be expressed recombinantly in bacteria and subsequently be analyzed. Hydroxylation of lysine residues changes the electrochemical properties of the collagen domains and could alter triple helix folding. The recombinant viral collagens could therefore be coexpressed with viral hydroxylases in order to verify the influence of lysyl hydroxylation. Many of the viral collagens further contain glycine rich repeats at the *C*-terminus. While similar stretches of glycine

occur in keratins and enable multimerization, expression of viral collagens with and without the *C*-terminal domains could reveal the involvement of glycine-rich repeats in triple helix assembly. In bacteria, collagen-like proteins occur in tail fibers and facilitate adhesion to host cells. A similar function of the viral collagens is plausible in infection of amoeba. With the CRISPR/Cas9 system adapted for expression in amoeba, collagen-like genes could be knocked out from the viral genome and the functional role of viral collagens could be investigated.

We used viral enzymes in order to stabilize recombinantly expressed human collagen. Recombinant collagens harbor a huge potential in biomedical applications due to the low immunogenicity and due to hemostatic properties. The viral prolyl 4-hydroxylase that we used to modify the recombinant human collagen does not possess the same sequence specificity as human prolyl 4-hydroxylase, which predominantly hydroxylates prolines at the Y-position in Gly-X-Y repeats. The effect of hydroxylation of proline residues at the X position on immunogenicity must be further evaluated. Many other giant viruses express prolyl 4-hydroxylases. It might hence be that these hydroxylases possess a different specificity than the mimivirus prolyl hydroxylase and would be more suitable for posttranslational modification of recombinant collagens. Our truncated collagen construct exhibits a melting temperature of 24°C. Full-length collagen needs to be expressed to ensure sufficient stability at body temperature. We included several steps of optimization such as bacterial codon usage and supplementation of the growth medium with glycine and proline, but failed to express full-length collagen type III in significant amounts. Optimization of other growth parameters such as different bacterial strains or growth in fed-batch culture are necessary to ensure successful expression of full-length collagen.

4. REFERENCES

1. Alberts, B., et al., *Molecular Biology of the Cell, Sixth Edition.* Molecular Biology of the Cell, Sixth Edition, 2015: p. 1-1342.
2. Lodish H, B.A., Zipursky SL, et al., *Molecular Cell Biology. 4th edition.* 2000: p. Section 22.3 Collagen: The Fibrous Proteins of the Matrix.
3. Kolble, K. and K.B. Reid, *The genomics of soluble proteins with collagenous domains: C1q, MBL, SP-A, SP-D, conglutinin, and CL-43.* Behring Inst Mitt, 1993(93): p. 81-6.
4. Gordon, M.K. and R.A. Hahn, *Collagens.* Cell Tissue Res, 2010. **339**(1): p. 247-57.
5. Ricard-Blum, S., *The collagen family.* Cold Spring Harb Perspect Biol, 2011. **3**(1): p. a004978.
6. Fratzl, P., *Collagen Structure and Mechanics.* 2008.
7. Exposito, J.Y., et al., *Evolution of collagens.* Anat Rec, 2002. **268**(3): p. 302-16.
8. Koch, M., et al., *Collagen XXIV, a vertebrate fibrillar collagen with structural features of invertebrate collagens: selective expression in developing cornea and bone.* J Biol Chem, 2003. **278**(44): p. 43236-44.
9. Boot-Handford, R.P., et al., *A novel and highly conserved collagen (pro(alpha)1(XXVII)) with a unique expression pattern and unusual molecular characteristics establishes a new clade within the vertebrate fibrillar collagen family.* J Biol Chem, 2003. **278**(33): p. 31067-77.
10. Harkness, M.L., R.D. Harkness, and D.W. James, *The effect of a protein-free diet on the collagen content of mice.* J Physiol, 1958. **144**(2): p. 307-13.
11. Liu, X., et al., *Type III collagen is crucial for collagen I fibrillogenesis and for normal cardiovascular development.* Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(5): p. 1852-1856.
12. Wenstrup, R.J., et al., *Type V collagen controls the initiation of collagen fibril assembly.* Journal of Biological Chemistry, 2004. **279**(51): p. 53331-53337.
13. Linsenmayer, T.F., et al., *Type V collagen: molecular structure and fibrillar organization of the chicken alpha 1(V) NH2-terminal domain, a putative regulator of corneal fibrillogenesis.* J Cell Biol, 1993. **121**(5): p. 1181-9.
14. Kramer, R.Z., et al., *Sequence dependent conformational variations of collagen triple-helical structure.* Nat Struct Biol, 1999. **6**(5): p. 454-7.
15. Emsley, J., et al., *Structural basis of collagen recognition by integrin alpha2beta1.* Cell, 2000. **101**(1): p. 47-56.
16. Sweeney, S.M., et al., *Defining the domains of type I collagen involved in heparin- binding and endothelial tube formation.* Proc Natl Acad Sci U S A, 1998. **95**(13): p. 7275-80.
17. Sweeney, S.M., et al., *Candidate cell and matrix interaction domains on the collagen fibril, the predominant protein of vertebrates.* J Biol Chem, 2008. **283**(30): p. 21187-97.
18. Beck, K., et al., *Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine.* Proc Natl Acad Sci U S A, 2000. **97**(8): p. 4273-8.
19. Bodian, D.L., et al., *Predicting the clinical lethality of osteogenesis imperfecta from collagen glycine mutations.* Biochemistry, 2008. **47**(19): p. 5424-32.
20. Ramshaw, J.A., N.K. Shah, and B. Brodsky, *Gly-X-Y tripeptide frequencies in collagen: a context for host-guest triple-helical peptides.* J Struct Biol, 1998. **122**(1-2): p. 86-91.
21. Kersteen, E.A. and R.T. Raines, *Contribution of tertiary amides to the conformational stability of collagen triple helices.* Biopolymers, 2001. **59**(1): p. 24-8.
22. Berg, R.A. and D.J. Prockop, *The thermal transition of a non-hydroxylated form of collagen. Evidence for a role for hydroxyproline in stabilizing the triple-helix of collagen.* Biochem Biophys Res Commun, 1973. **52**(1): p. 115-20.
23. Sakakibara, S., et al., *Synthesis of (Pro-Hyp-Gly) n of defined molecular weights. Evidence for the stabilization of collagen triple helix by hydroxypyroline.* Biochim Biophys Acta, 1973. **303**(1): p. 198-202.

24. Exposito, J.Y., et al., *The fibrillar collagen family.* Int J Mol Sci, 2010. **11**(2): p. 407-26.
25. Bourhis, J.M., et al., *Structural basis of fibrillar collagen trimerization and related genetic disorders.* Nat Struct Mol Biol, 2012. **19**(10): p. 1031-6.
26. Mazzorana, M., et al., *Mechanisms of collagen trimer formation. Construction and expression of a recombinant minigene in HeLa cells reveals a direct effect of prolyl hydroxylation on chain assembly of type XII collagen.* J Biol Chem, 1993. **268**(5): p. 3029-32.
27. Woodley, D.T., et al., *Collagen Telopeptides (Cross-Linking Sites) Play a Role in Collagen Gel Lattice Contraction.* Journal of Investigative Dermatology, 1991. **97**(3): p. 580-585.
28. Gelse, K., E. Poschl, and T. Aigner, *Collagens - structure, function, and biosynthesis.* Advanced Drug Delivery Reviews, 2003. **55**(12): p. 1531-1546.
29. Yada, T., et al., *Occurrence in Chick-Embryo Vitreous-Humor of a Type-Ix Collagen Proteoglycan with an Extraordinarily Large Chondroitin Sulfate Chain and Short Alpha-1 Polypeptide.* Journal of Biological Chemistry, 1990. **265**(12): p. 6992-6999.
30. Shaw, L.M. and B.R. Olsen, *FACIT collagens: diverse molecular bridges in extracellular matrices.* Trends Biochem Sci, 1991. **16**(5): p. 191-4.
31. Weber, S., et al., *Identification of 47 novel mutations in patients with Alport syndrome and thin basement membrane nephropathy.* Pediatr Nephrol, 2016.
32. Bruckner-Tuderman, L. and C. Has, *Disorders of the cutaneous basement membrane zone--the paradigm of epidermolysis bullosa.* Matrix Biol, 2014. **33**: p. 29-34.
33. Myllyharju, J. and K.I. Kivirikko, *Collagens, modifying enzymes and their mutations in humans, flies and worms.* Trends Genet, 2004. **20**(1): p. 33-43.
34. Ferreira, L.R., et al., *Association of Hsp47, Grp78, and Grp94 with procollagen supports the successive or coupled action of molecular chaperones.* J Cell Biochem, 1994. **56**(4): p. 518-26.
35. Malhotra, V. and P. Erlmann, *Protein export at the ER: loading big collagens into COPII carriers.* EMBO J, 2011. **30**(17): p. 3475-80.
36. Jin, L., et al., *Ubiquitin-dependent regulation of COPII coat size and function.* Nature, 2012. **482**(7386): p. 495-500.
37. Stephens, D.J., *Cell biology: Collagen secretion explained.* Nature, 2012. **482**(7386): p. 474-5.
38. Malhotra, V. and P. Erlmann, *The Pathway of Collagen Secretion.* Annu Rev Cell Dev Biol, 2015. **31**: p. 109-24.
39. Tang, B.L. and W. Hong, *ADAMTS: a novel family of proteases with an ADAM protease domain and thrombospondin 1 repeats.* FEBS Lett, 1999. **445**(2-3): p. 223-5.
40. Prockop, D.J., A.L. Sieron, and S.W. Li, *Procollagen N-proteinase and procollagen C-proteinase. Two unusual metalloproteinases that are essential for procollagen processing probably have important roles in development and cell signaling.* Matrix Biol, 1998. **16**(7): p. 399-408.
41. Kadler, K.E., Y. Hojima, and D.J. Prockop, *Assembly of collagen fibrils de novo by cleavage of the type I pC-collagen with procollagen C-proteinase. Assay of critical concentration demonstrates that collagen self-assembly is a classical example of an entropy-driven process.* J Biol Chem, 1987. **262**(32): p. 15696-701.
42. Lucero, H.A. and H.M. Kagan, *Lysyl oxidase: an oxidative enzyme and effector of cell function.* Cell Mol Life Sci, 2006. **63**(19-20): p. 2304-16.
43. Holmgren, S.K., et al., *A hyperstable collagen mimic.* Chem Biol, 1999. **6**(2): p. 63-70.
44. Prockop, D.J. and K.I. Kivirikko, *Effect of polymer size on the inhibition of protocollagen proline hydroxylase by polyproline II.* J Biol Chem, 1969. **244**(18): p. 4838-42.
45. Bhatnagar, R.S., R.S. Rapaka, and D.W. Urry, *Interaction of polypeptide models of elastin with prolyl hydroxylase.* FEBS Lett, 1978. **95**(1): p. 61-4.
46. Rauch, F., et al., *Cole-Carpenter Syndrome Is Caused by a Heterozygous Missense Mutation in P4HB.* American Journal of Human Genetics, 2015. **96**(3): p. 425-431.
47. Rhodes, R.K. and E.J. Miller, *Physicochemical characterization and molecular organization of the collagen A and B chains.* Biochemistry, 1978. **17**(17): p. 3442-8.
48. Gryder, R.M., M. Lamon, and E. Adams, *Sequence position of 3-hydroxyproline in basement membrane collagen. Isolation of glycyl-3-hydroxyprolyl-4-hydroxyproline from swine kidney.* J Biol Chem, 1975. **250**(7): p. 2470-4.

49. Mizuno, K., et al., *Effect of the -Gly-3(S)-hydroxyprolyl-4(R)-hydroxyprolyl- tripeptide unit on the stability of collagen model peptides.* FEBS J, 2008. **275**(23): p. 5830-40.
50. Cabral, W.A., et al., *Prolyl 3-hydroxylase 1 deficiency causes a recessive metabolic bone disorder resembling lethal/severe osteogenesis imperfecta.* Nat Genet, 2007. **39**(3): p. 359-65.
51. Elsas, L.J., 2nd, R.L. Miller, and S.R. Pinnell, *Inherited human collagen lysyl hydroxylase deficiency: ascorbic acid response.* J Pediatr, 1978. **92**(3): p. 378-84.
52. Myllyla, R., et al., *Expanding the lysyl hydroxylase toolbox: new insights into the localization and activities of lysyl hydroxylase 3 (LH3).* J Cell Physiol, 2007. **212**(2): p. 323-9.
53. Wang, C., M. Valtavaara, and R. Myllyla, *Lack of collagen type specificity for lysyl hydroxylase isoforms.* DNA Cell Biol, 2000. **19**(2): p. 71-7.
54. Bank, R.A., et al., *Defective collagen crosslinking in bone, but not in ligament or cartilage, in Bruck syndrome: Indications for a bone-specific telopeptide lysyl hydroxylase on chromosome 17.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(3): p. 1054-1058.
55. van der Slot, A.J., et al., *Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis.* Journal of Biological Chemistry, 2003. **278**(42): p. 40967-40972.
56. Giunta, C., et al., *Nevo syndrome is allelic to the kyphoscoliotic type of the Ehlers-Danlos syndrome (EDS VIA).* American Journal of Medical Genetics Part A, 2005. **133a**(2): p. 158-164.
57. Puig-Hervas, M.T., et al., *Mutations in PLOD2 Cause Autosomal-Recessive Connective Tissue Disorders Within the Bruck Syndrome-Osteogenesis Imperfecta Phenotypic Spectrum.* Human Mutation, 2012. **33**(10): p. 1444-1449.
58. Salo, A.M., et al., *A Connective Tissue Disorder Caused by Mutations of the Lysyl Hydroxylase 3 Gene.* American Journal of Human Genetics, 2008. **83**(4): p. 495-503.
59. Spiro, R.G., *The structure of the disaccharide unit of the renal glomerular basement membrane.* J Biol Chem, 1967. **242**(20): p. 4813-23.
60. Schegg, B., et al., *Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases.* Mol Cell Biol, 2009. **29**(4): p. 943-52.
61. Liefhebber, J.M., et al., *The human collagen beta(1-O)galactosyltransferase, GLT25D1, is a soluble endoplasmic reticulum localized protein.* BMC Cell Biol, 2010. **11**: p. 33.
62. Spiro, M.J. and R.G. Spiro, *Studies on the biosynthesis of the hydroxylsine-linked disaccharide unit of basement membranes and collagens. II. Kidney galactosyltransferase.* J Biol Chem, 1971. **246**(16): p. 4910-8.
63. Ruotsalainen, H., et al., *Glycosylation catalyzed by lysyl hydroxylase 3 is essential for basement membranes.* J Cell Sci, 2006. **119**(Pt 4): p. 625-35.
64. Brownell, A.G. and A. Veis, *The intracellular location of the glycosylation of hydroxylysine of collagen.* Biochem Biophys Res Commun, 1975. **63**(2): p. 371-7.
65. Harwood, R., M.E. Grant, and D.S. Jackson, *Studies on the glycosylation of hydroxylysine residues during collagen biosynthesis and the subcellular localization of collagen galactosyltransferase and collagen glucosyltransferase in tendon and cartilage cells.* Biochem J, 1975. **152**(2): p. 291-302.
66. Kivirikko, K.I. and R. Myllyla, *Collagen glycosyltransferases.* Int Rev Connect Tissue Res, 1979. **8**: p. 23-72.
67. Sipila, L., et al., *Secretion and assembly of type IV and VI collagens depend on glycosylation of hydroxylysines.* J Biol Chem, 2007. **282**(46): p. 33381-8.
68. Norman, K.R. and D.G. Moerman, *The let-268 locus of Caenorhabditis elegans encodes a procollagen lysyl hydroxylase that is essential for type IV collagen secretion.* Dev Biol, 2000. **227**(2): p. 690-705.
69. Engelholm, L.H., et al., *The urokinase receptor associated protein (uPARAP/endo180): a novel internalization receptor connected to the plasminogen activation system.* Trends Cardiovasc Med, 2001. **11**(1): p. 7-13.
70. Jurgensen, H.J., et al., *A novel functional role of collagen glycosylation: interaction with the endocytic collagen receptor uparap/ENDO180.* J Biol Chem, 2011. **286**(37): p. 32736-48.

71. Engelholm, L.H., et al., *uPARAP/Endo180 is essential for cellular uptake of collagen and promotes fibroblast collagen adhesion.* J Cell Biol, 2003. **160**(7): p. 1009-15.
72. Stawikowski, M.J., et al., *Glycosylation modulates melanoma cell alpha2beta1 and alpha3beta1 integrin interactions with type IV collagen.* J Biol Chem, 2014. **289**(31): p. 21591-604.
73. Terajima, M., et al., *Glycosylation and cross-linking in bone type I collagen.* J Biol Chem, 2014. **289**(33): p. 22636-47.
74. Pokidysheva, E., et al., *Posttranslational modifications in type I collagen from different tissues extracted from wild type and prolyl 3-hydroxylase 1 null mice.* J Biol Chem, 2013. **288**(34): p. 24742-52.
75. McPherson, J.D., B.H. Shilton, and D.J. Walton, *Role of fructose in glycation and cross-linking of proteins.* Biochemistry, 1988. **27**(6): p. 1901-7.
76. Lalla, E., et al., *Hyperglycemia, glycoxidation and receptor for advanced glycation endproducts: potential mechanisms underlying diabetic complications, including diabetes-associated periodontitis.* Periodontol 2000, 2000. **23**: p. 50-62.
77. Ahmed, N., *Advanced glycation endproducts--role in pathology of diabetic complications.* Diabetes Res Clin Pract, 2005. **67**(1): p. 3-21.
78. Saito, M. and K. Marumo, *Collagen cross-links as a determinant of bone quality: a possible explanation for bone fragility in aging, osteoporosis, and diabetes mellitus.* Osteoporos Int, 2010. **21**(2): p. 195-214.
79. Bruel, A. and H. Oxlund, *Changes in biomechanical properties, composition of collagen and elastin, and advanced glycation endproducts of the rat aorta in relation to age.* Atherosclerosis, 1996. **127**(2): p. 155-65.
80. Yan, S.F., R. Ramasamy, and A.M. Schmidt, *Mechanisms of disease: advanced glycation end-products and their receptor in inflammation and diabetes complications.* Nat Clin Pract Endocrinol Metab, 2008. **4**(5): p. 285-93.
81. Bierhaus, A., et al., *Diabetes-associated sustained activation of the transcription factor nuclear factor-kappaB.* Diabetes, 2001. **50**(12): p. 2792-808.
82. Oleniuc, M., et al., *Consequences of Advanced Glycation End Products Accumulation in Chronic Kidney Disease and Clinical Usefulness of Their Assessment Using a Non-invasive Technique - Skin Autofluorescence.* Maedica (Buchar), 2011. **6**(4): p. 298-307.
83. Bodiga, V.L., S.R. Eda, and S. Bodiga, *Advanced glycation end products: role in pathology of diabetic cardiomyopathy.* Heart Fail Rev, 2014. **19**(1): p. 49-63.
84. Stitt, A.W., *Advanced glycation: an important pathological event in diabetic and age related ocular disease.* Br J Ophthalmol, 2001. **85**(6): p. 746-53.
85. Siegel, R.C., *Biosynthesis of collagen crosslinks: increased activity of purified lysyl oxidase with reconstituted collagen fibrils.* Proc Natl Acad Sci U S A, 1974. **71**(12): p. 4826-30.
86. Knott, L. and A.J. Bailey, *Collagen cross-links in mineralizing tissues: a review of their chemistry, function, and clinical relevance.* Bone, 1998. **22**(3): p. 181-7.
87. Pez, F., et al., *The HIF-1-inducible lysyl oxidase activates HIF-1 via the Akt pathway in a positive regulation loop and synergizes with HIF-1 in promoting tumor cell growth.* Cancer Res, 2011. **71**(5): p. 1647-57.
88. Denko, N.C., et al., *Investigating hypoxic tumor physiology through gene expression patterns.* Oncogene, 2003. **22**(37): p. 5907-14.
89. Gilkes, D.M., et al., *Hypoxia-inducible factor 1 (HIF-1) promotes extracellular matrix remodeling under hypoxic conditions by inducing P4HA1, P4HA2, and PLOD2 expression in fibroblasts.* J Biol Chem, 2013. **288**(15): p. 10819-29.
90. Erler, J.T., et al., *Lysyl oxidase is essential for hypoxia-induced metastasis.* Nature, 2006. **440**(7088): p. 1222-6.
91. Mark Brewer, M.H., *Collagen Solutions plc Initiation.* hardman&co, 2014.
92. Mullins, R.J., C. Richards, and T. Walker, *Allergic reactions to oral, surgical and topical bovine collagen - Anaphylactic risk for surgeons.* Australian and New Zealand Journal of Ophthalmology, 1996. **24**(3): p. 257-260.

93. Freyman, T.M., et al., *Fibroblast contraction of a collagen-GAG matrix.* Biomaterials, 2001. **22**(21): p. 2883-2891.
94. Gorham, S.D., et al., *Cellular invasion and breakdown of three different collagen films in the lumbar muscle of the rat.* Biomaterials, 1990. **11**(2): p. 113-8.
95. Chaudhary, S., et al., *Use of gentamicin-loaded collagen sponge in internal fixation of open fractures.* Chin J Traumatol, 2011. **14**(4): p. 209-14.
96. Choi, Y., et al., *Sinus augmentation using absorbable collagen sponge loaded with Escherichia coli-expressed recombinant human bone morphogenetic protein 2 in a standardized rabbit sinus model: a radiographic and histologic analysis.* Clin Oral Implants Res, 2012. **23**(6): p. 682-9.
97. Swieringa, A.J., et al., *In vivo pharmacokinetics of a gentamicin-loaded collagen sponge in acute periprosthetic infection: serum values in 19 patients.* Acta Orthop, 2008. **79**(5): p. 637-42.
98. Ruszczak, Z. and W. Friess, *Collagen as a carrier for on-site delivery of antibacterial drugs.* Advanced Drug Delivery Reviews, 2003. **55**(12): p. 1679-1698.
99. Papi, M., et al., *Controlled self assembly of collagen nanoparticle.* Journal of Nanoparticle Research, 2011. **13**(11): p. 6141-6147.
100. Varani, J., et al., *Decreased collagen production in chronologically aged skin - Roles of age-dependent alteration in fibroblast function and defective mechanical stimulation.* American Journal of Pathology, 2006. **168**(6): p. 1861-1868.
101. Sandoval, L., et al., *Trends in the use of neurotoxins and dermal fillers by US physicians.* Journal of the American Academy of Dermatology, 2015. **72**(5): p. Ab21-Ab21.
102. Yu, Z.X., et al., *Bacterial collagen-like proteins that form triple-helical structures.* Journal of Structural Biology, 2014. **186**(3): p. 451-461.
103. Yu, Z.X., B. Brodsky, and M. Inouye, *Dissecting a Bacterial Collagen Domain from Streptococcus pyogenes SEQUENCE AND LENGTH-DEPENDENT VARIATIONS IN TRIPLE HELIX STABILITY AND FOLDING.* Journal of Biological Chemistry, 2011. **286**(21): p. 18960-18968.
104. Peng, Y.Y., et al., *Towards scalable production of a collagen-like protein from Streptococcus pyogenes for biomedical applications.* Microbial Cell Factories, 2012. **11**.
105. Peng, Y.Y., et al., *A Streptococcus pyogenes derived collagen-like protein as a non-cytotoxic and non-immunogenic cross-linkable biomaterial.* Biomaterials, 2010. **31**(10): p. 2755-2761.
106. Guo, J.Q., et al., *Medium optimization based on the metabolic-flux spectrum of recombinant Escherichia coli for high expression of human-like collagen II.* Biotechnology and Applied Biochemistry, 2010. **57**: p. 55-62.
107. Neubauer, A., P. Neubauer, and J. Myllyharju, *High-level production of human collagen prolyl 4-hydroxylase in Escherichia coli.* Matrix Biol, 2005. **24**(1): p. 59-68.
108. Pinkas, D.M., et al., *Tunable, post-translational hydroxylation of collagen Domains in Escherichia coli.* ACS Chem Biol, 2011. **6**(4): p. 320-4.
109. Buechter, D.D., et al., *Co-translational incorporation of trans-4-hydroxyproline into recombinant proteins in bacteria.* J Biol Chem, 2003. **278**(1): p. 645-50.
110. Myllyharju, J., et al., *Expression of recombinant human type I-III collagens in the yeast pichia pastoris.* Biochem Soc Trans, 2000. **28**(4): p. 353-7.
111. Keizer-Gunnink, I., et al., *Accumulation of properly folded human type III procollagen molecules in specific intracellular membranous compartments in the yeast Pichia pastoris.* Matrix Biol, 2000. **19**(1): p. 29-36.
112. Pakkanen, O., et al., *Assembly of stable human type I and III collagen molecules from hydroxylated recombinant chains in the yeast Pichia pastoris. Effect of an engineered C-terminal oligomerization domain foldon.* J Biol Chem, 2003. **278**(34): p. 32478-83.
113. Polarek, J.W., et al., *Production and uses of recombinant human collagen and derivatives.* Abstracts of Papers of the American Chemical Society, 2004. **227**: p. U310-U310.
114. Yang, C., et al., *Development of a recombinant human collagen-type III based hemostat.* Journal of Biomedical Materials Research Part B-Applied Biomaterials, 2004. **69b**(1): p. 18-24.
115. Yang, C.L., et al., *The application of recombinant human collagen in tissue engineering.* Biodrugs, 2004. **18**(2): p. 103-119.

116. Toman, P.D., et al., *Production of recombinant human type I procollagen trimers using a four-gene expression system in the yeast Saccharomyces cerevisiae.* Journal of Biological Chemistry, 2000. **275**(30): p. 23303-23309.
117. Olsen, D.R., et al., *Production of human type I collagen in yeast reveals unexpected new insights into the molecular assembly of collagen trimers.* Journal of Biological Chemistry, 2001. **276**(26): p. 24038-24043.
118. Barta, A., et al., *The Expression of a Nopaline Synthase - Human Growth-Hormone Chimeric Gene in Transformed Tobacco and Sunflower Callus-Tissue.* Plant Molecular Biology, 1986. **6**(5): p. 347-357.
119. Fischer, R., Y.C. Liao, and J. Drossard, *Affinity-purification of a TMV-specific recombinant full-size antibody from a transgenic tobacco suspension culture.* Journal of Immunological Methods, 1999. **226**(1-2): p. 1-10.
120. Ruggiero, F., et al., *Triple helix assembly and processing of human collagen produced in transgenic tobacco plants.* Febs Letters, 2000. **469**(1): p. 132-136.
121. Tanaka, M., K. Sato, and T. Uchida, *Plant Prolyl Hydroxylase Recognizes Poly(L-Proline) Ii Helix.* Journal of Biological Chemistry, 1981. **256**(22): p. 1397-1400.
122. Bosch, D. and A. Schots, *Plant glycans: friend or foe in vaccine development?* Expert Review of Vaccines, 2010. **9**(8): p. 835-842.
123. Stein, H., et al., *Production of Bioactive, Post-Translationally Modified, Heterotrimeric, Human Recombinant Type-I Collagen in Transgenic Tobacco.* Biomacromolecules, 2009. **10**(9): p. 2640-2645.
124. Jarvis, D.L., *Baculovirus-Insect Cell Expression Systems.* Guide to Protein Purification, Second Edition, 2009. **463**: p. 191-222.
125. Fertala, A., et al., *Self-Assembly into Fibrils of Collagen-Ii by Enzymatic Cleavage of Recombinant Procollagen-Ii - Lag Period, Critical Concentration, and Morphology of Fibrils Differ from Collagen-I.* Journal of Biological Chemistry, 1994. **269**(15): p. 11584-11589.
126. Geddis, A.E. and D.J. Prockop, *Expression of Human Col1a1 Gene in Stably Transfected Ht-1080 Cells - the Production of a Thermostable Homotrimer of Type-1 Collagen in a Recombinant System.* Matrix, 1993. **13**(5): p. 399-405.
127. Fichard, A., et al., *Human recombinant alpha 1(V) collagen chain - Homotrimeric assembly and subsequent processing.* Journal of Biological Chemistry, 1997. **272**(48): p. 30083-30087.
128. Chen, M., et al., *The recombinant expression of full-length type VII collagen and characterization of molecular mechanisms underlying dystrophic epidermolysis bullosa.* Journal of Biological Chemistry, 2002. **277**(3): p. 2118-2124.
129. Hou, Y.P., et al., *Intravenously Administered Recombinant Human Type VII Collagen Derived from Chinese Hamster Ovary Cells Reverses the Disease Phenotype in Recessive Dystrophic Epidermolysis Bullosa Mice.* Journal of Investigative Dermatology, 2015. **135**(12): p. 3060-3067.
130. Toman, P.D., et al., *Production of recombinant human type I procollagen homotrimer in the mammary gland of transgenic mice.* Transgenic Research, 1999. **8**(6): p. 415-427.
131. Rasmussen, M., M. Jacobsson, and L. Bjorck, *Genome-based identification and analysis of collagen-related structural motifs in bacterial and viral proteins.* Journal of Biological Chemistry, 2003. **278**(34): p. 32313-32316.
132. Bann, J.G., D.H. Peyton, and H.P. Bachinger, *Sweet is stable: glycosylation stabilizes collagen.* Febs Letters, 2000. **473**(2): p. 237-240.
133. Mann, K., et al., *Glycosylated threonine but not 4-hydroxyproline dominates the triple helix stabilizing positions in the sequence of a hydrothermal vent worm cuticle collagen.* Journal of Molecular Biology, 1996. **261**(2): p. 255-266.
134. Boydston, J.A., et al., *Orientation within the exosporium and structural stability of the collagen-like glycoprotein BclA of Bacillus anthracis.* Journal of Bacteriology, 2005. **187**(15): p. 5310-5317.
135. Xu, C.Y., et al., *Expanding the Family of Collagen Proteins: Recombinant Bacterial Collagens of Varying Composition Form Triple-Helices of Similar Stability.* Biomacromolecules, 2010. **11**(2): p. 348-356.

136. Han, R.L., et al., *Binding of the low-density lipoprotein by streptococcal collagen-like protein Scl1 of Streptococcus pyogenes.* Molecular Microbiology, 2006. **61**(2): p. 351-367.
137. Gao, Y.M., et al., *The Scl1 of M41-type group A Streptococcus binds the high-density lipoprotein.* Fems Microbiology Letters, 2010. **309**(1): p. 55-61.
138. Caswell, C.C., et al., *The Scl1 protein of M6-type group A Streptococcus binds the human complement regulatory protein, factor H, and inhibits the alternative pathway of complement.* Molecular Microbiology, 2008. **67**(3): p. 584-596.
139. Reuter, M., et al., *Binding of the Human Complement Regulators CFHR1 and Factor H by Streptococcal Collagen-like Protein 1 (Scl1) via Their Conserved C Termini Allows Control of the Complement Cascade at Multiple Levels.* Journal of Biological Chemistry, 2010. **285**(49): p. 38473-38485.
140. Caswell, C.C., et al., *Identification of the First Prokaryotic Collagen Sequence Motif That Mediates Binding to Human Collagen Receptors, Integrins alpha(2)beta(1) and alpha(11)beta(1).* Journal of Biological Chemistry, 2008. **283**(52): p. 36168-36175.
141. Humtsoe, J.O., et al., *A streptococcal collagen-like protein interacts with the alpha(2)beta(1) integrin and induces intracellular signaling.* Journal of Biological Chemistry, 2005. **280**(14): p. 13848-13857.
142. Smith, M.C.M., et al., *Bacteriophage collagen.* Science, 1998. **279**(5358): p. 1834-1834.
143. Ghosh, N., et al., *Collagen-Like Proteins in Pathogenic E. coli Strains.* Plos One, 2012. **7**(6).
144. Haggardljungquist, E., C. Halling, and R. Calendar, *DNA-Sequences of the Tail Fiber Genes of Bacteriophage P2 - Evidence for Horizontal Transfer of Tail Fiber Genes among Unrelated Bacteriophages.* Journal of Bacteriology, 1992. **174**(5): p. 1462-1477.
145. Zairi, M., et al., *The Collagen-like Protein gp12 Is a Temperature-dependent Reversible Binder of SPP1 Viral Capsids.* Journal of Biological Chemistry, 2014. **289**(39): p. 27169-27181.
146. Vanetten, J.L., et al., *Virus-Infection of Culturable Chlorella-Like Algae and Development of a Plaque-Assay.* Science, 1983. **219**(4587): p. 994-996.
147. Suzan-Monti, M., B. La Scola, and D. Raoult, *Genomic and evolutionary aspects of Mimivirus.* Virus Research, 2006. **117**(1): p. 145-155.
148. Yutin, N., et al., *Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution.* Virology Journal, 2009. **6**.
149. Koonin, E.V. and N. Yutin, *Origin and Evolution of Eukaryotic Large Nucleo-Cytoplasmic DNA Viruses.* Intervirology, 2010. **53**(5): p. 284-292.
150. Yutin, N., Y.I. Wolf, and E.V. Koonin, *Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life.* Virology, 2014. **466-467**: p. 38-52.
151. Luther, K.B., et al., *Mimivirus Collagen Is Modified by Bifunctional Lysyl Hydroxylase and Glycosyltransferase Enzyme.* Journal of Biological Chemistry, 2011. **286**(51): p. 43701-43709.
152. Shah, N., et al., *Exposure to Mimivirus Collagen Promotes Arthritis.* Journal of Virology, 2014. **88**(2): p. 838-845.
153. Rutschmann, C., et al., *Recombinant expression of hydroxylated human collagen in Escherichia coli.* Applied Microbiology and Biotechnology, 2014. **98**(10): p. 4445-4455.
154. La Scola, B., et al., *The virophage as a unique parasite of the giant mimivirus.* Nature, 2008. **455**(7209): p. 100-U65.
155. Desnues, C., et al., *Provirophages and transpovirons as the diverse mobilome of giant viruses.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(44): p. 18078-18083.
156. Gaia, M., et al., *Broad Spectrum of Mimiviridae Virophage Allows Its Isolation Using a Mimivirus Reporter.* Plos One, 2013. **8**(4).
157. Gaia, M., et al., *Zamilon, a Novel Virophage with Mimiviridae Host Specificity.* Plos One, 2014. **9**(4).
158. Yau, S., et al., *Virophage control of antarctic algal host-virus dynamics.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(15): p. 6163-6168.
159. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Research, 2014. **42**(D1): p. D222-D230.

160. Thompson, B.M., et al., *Assembly of the BclB glycoprotein into the exosporium and evidence for its role in the formation of the exosporium cap' structure in Bacillus anthracis.* Molecular Microbiology, 2012. **86**(5): p. 1073-1084.
161. Russ, W.P. and D.M. Engelman, *The GxxxG motif: a framework for transmembrane helix-helix association.* J Mol Biol, 2000. **296**(3): p. 911-9.
162. Wilson, A.K., P.A. Coulombe, and E. Fuchs, *The Roles of K5 and K14 Head, Tail, and R/K L L E G E Domains in Keratin Filament Assembly Invitro.* Journal of Cell Biology, 1992. **119**(2): p. 401-414.
163. Lee, C.H., et al., *Structural basis for heteromeric assembly and perinuclear organization of keratin filaments.* Nature Structural & Molecular Biology, 2012. **19**(7): p. 707-+.
164. Longbotham, J.E., et al., *Structure and Mechanism of a Viral Collagen Prolyl Hydroxylase.* Biochemistry, 2015. **54**(39): p. 6093-6105.
165. Gorres, K.L. and R.T. Raines, *Prolyl 4-hydroxylase.* Crit Rev Biochem Mol Biol, 2010. **45**(2): p. 106-24.
166. Szpak, P., *Fish bone chemistry and ultrastructure: implications for taphonomy and stable isotope analysis.* Journal of Archaeological Science, 2011. **38**(12): p. 3358-3372.
167. Mohs, A., et al., *Mechanism of stabilization of a bacterial collagen triple helix in the absence of hydroxyproline.* Journal of Biological Chemistry, 2007. **282**(41): p. 29757-29765.
168. Fallas, J.A., et al., *Structural Insights into Charge Pair Interactions in Triple Helical Collagen-like Proteins.* Journal of Biological Chemistry, 2012. **287**(11): p. 8039-8047.
169. Persikov, A.V., et al., *Electrostatic interactions involving lysine make major contributions to collagen triple-helix stability.* Biochemistry, 2005. **44**(5): p. 1414-1422.
170. Moreira, D. and C. Brochier-Armanet, *Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes.* BMC Evol Biol, 2008. **8**: p. 12.
171. Filee, J., *Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution.* Front Microbiol, 2015. **6**: p. 593.
172. Claverie, J.M. and C. Abergel, *Mimivirus and its Virophage.* Annual Review of Genetics, 2009. **43**: p. 49-66.
173. Boyer, M., et al., *Mimivirus shows dramatic genome reduction after intraamoebal culture.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(25): p. 10296-10301.
174. Goldstein, A. and E. Adams, *Glycylhydroxyprolyl sequences in earthworm cuticle collagen: glycylhydroxyprolylserine.* J Biol Chem, 1970. **245**(20): p. 5478-83.
175. Inouye, K., et al., *Synthesis and physical properties of (hydroxyproline-proline-glycine)10: hydroxyproline in the X-position decreases the melting temperature of the collagen triple helix.* Arch Biochem Biophys, 1982. **219**(1): p. 198-203.
176. Grassmann, W. and H. Schleich, *On the carbo-hydrate-content of collagen II Announcement on knowledge of collagen.* Biochemische Zeitschrift, 1935. **277**: p. 320-328.
177. Schegg, B., et al., *Core Glycosylation of Collagen Is Initiated by Two beta(1-O)Galactosyltransferases.* Molecular and Cellular Biology, 2009. **29**(4): p. 943-952.
178. Ruotsalainen, H., et al., *Glycosylation catalyzed by lysyl hydroxylase 3 is essential for basement membranes.* Journal of Cell Science, 2006. **119**(4): p. 625-635.
179. Janik, M.E., A. Litynska, and P. Vereecken, *Cell migration-The role of integrin glycosylation.* Biochimica Et Biophysica Acta-General Subjects, 2010. **1800**(6): p. 545-555.
180. Varki, A., *Selectin Ligands.* Proceedings of the National Academy of Sciences of the United States of America, 1994. **91**(16): p. 7390-7397.
181. Schwab, I. and F. Nimmerjahn, *Role of sialylation in the anti-inflammatory activity of intravenous immunoglobulin - F(ab ')(2) versus Fc sialylation.* Clinical and Experimental Immunology, 2014. **178**: p. 97-99.
182. Ghosh, A.K., *Factors involved in the regulation of type I collagen gene expression: Implication in fibrosis.* Experimental Biology and Medicine, 2002. **227**(5): p. 301-314.
183. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nature Genetics, 2003. **33**: p. 245-254.

184. Delcuve, G.P., M. Rastegar, and J.R. Davie, *Epigenetic Control.* Journal of Cellular Physiology, 2009. **219**(2): p. 243-250.
185. Illingworth, R.S. and A.P. Bird, *CpG islands - 'A rough guide'.* Febs Letters, 2009. **583**(11): p. 1713-1720.
186. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends in Genetics, 1997. **13**(8): p. 335-340.
187. Baylin, S.B., et al., *Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer.* Human Molecular Genetics, 2001. **10**(7): p. 687-692.
188. Vincent, Z.L., M.D. Mitchell, and A.P. Ponnampalam, *Epigenetic Regulation of Matrix Metalloproteinases and Their Inhibitors in Parturition.* Reproductive Sciences, 2012. **19**(S3): p. 188a-188a.
189. Ha, V.T., et al., *A Patient with Ehlers-Danlos Syndrome Type-Vi Is a Compound Heterozygote for Mutations in the Lysyl Hydroxylase Gene.* Journal of Clinical Investigation, 1994. **93**(4): p. 1716-1721.
190. Ha-Vinh, R., et al., *Phenotypic and molecular characterization of Bruck syndrome (osteogenesis imperfecta with contractures of the large joints) caused by a recessive mutation in PLOD2.* American Journal of Medical Genetics Part A, 2004. **131a**(2): p. 115-120.
191. Norman, K.R. and D.G. Moerman, *The let-268 locus of Caenorhabditis elegans encodes a procollagen lysyl hydroxylase that is essential for type IV collagen secretion.* Developmental Biology, 2000. **227**(2): p. 690-705.
192. Kamath, R.S., et al., *Systematic functional analysis of the Caenorhabditis elegans genome using RNAi.* Nature, 2003. **421**(6920): p. 231-237.
193. Chernousov, M.A., R.C. Stahl, and D.J. Carey, *Schwann cell type v collagen inhibits axonal outgrowth and promotes Schwann cell migration via distinct adhesive activities of the collagen and noncollagen domains.* Journal of Neuroscience, 2001. **21**(16): p. 6125-6135.
194. Boot-Handford, R.P. and D.S. Tuckwell, *Fibrillar collagen: the key to vertebrate evolution? A tale of molecular incest.* Bioessays, 2003. **25**(2): p. 142-51.
195. Yamada, Y., et al., *The Collagen Gene - Evidence for Its Evolutionary Assembly by Amplification of a DNA Segment Containing an Exon of 54 Bp.* Cell, 1980. **22**(3): p. 887-892.
196. Exposito, J.Y., M. Vanderrest, and R. Garrone, *The Complete Intron-Exon Structure of Ephydatia-Mulleri Fibrillar Collagen Gene Suggests a Mechanism for the Evolution of an Ancestral Gene Module.* Journal of Molecular Evolution, 1993. **37**(3): p. 254-259.
197. Exposito, J.Y., et al., *Evolution of collagens.* Anatomical Record, 2002. **268**(3): p. 302-316.
198. Klose, T., et al., *The three-dimensional structure of Mimivirus.* Intervirology, 2010. **53**(5): p. 268-73.
199. Xiao, J., et al., *Local conformation and dynamics of isoleucine in the collagenase cleavage site provide a recognition signal for matrix metalloproteinases.* J Biol Chem, 2010. **285**(44): p. 34181-90.
200. *The American Society for Aesthetic Plastic Surgery's Cosmetic Surgery National Data Bank: Statistics 2013.* Aesthetic Surgery Journal, 2014. **34**: p. 1-20.
201. Richter, A.W., E.M. Ryde, and E.O. Zetterstrom, *Non-Immunogenicity of a Purified Sodium Hyaluronate Preparation in Man.* International Archives of Allergy and Applied Immunology, 1979. **59**(1): p. 45-48.
202. Honig, J.F., U. Brink, and M. Korabiowska, *Severe granulomatous allergic tissue reaction after hyaluronic acid injection in the treatment of facial lines and its surgical correction.* Journal of Craniofacial Surgery, 2003. **14**(2): p. 197-200.
203. Holmstro.B and J. Ricica, *Production of Hyaluronic Acid by a Streptococcal Strain in Batch Culture.* Applied Microbiology, 1967. **15**(6): p. 1409-&.
204. Laurent, T.C., *Structure and Function of Hyaluronic Acid.* Scandinavian Journal of Clinical & Laboratory Investigation, 1969. **S 23**: p. 10-&.
205. Smith, K.C., *New fillers for the new man.* Dermatologic Therapy, 2007. **20**(6): p. 388-393.
206. Verzijl, N., et al., *Effect of collagen turnover on the accumulation of advanced glycation end products.* Journal of Biological Chemistry, 2000. **275**(50): p. 39027-39031.

207. Brown, T.J., U.B.G. Laurent, and J.R.E. Fraser, *Turnover of Hyaluronan in Synovial Joints - Elimination of Labeled Hyaluronan from the Knee-Joint of the Rabbit.* Experimental Physiology, 1991. **76**(1): p. 125-134.
208. Tomizawa, Y., *Clinical benefits and risk analysis of topical hemostats: a review.* J Artif Organs, 2005. **8**(3): p. 137-42.
209. Haugh, M.G., et al., *Crosslinking and Mechanical Properties Significantly Influence Cell Attachment, Proliferation, and Migration Within Collagen Glycosaminoglycan Scaffolds.* Tissue Engineering Part A, 2011. **17**(9-10): p. 1201-1208.

5. ACKNOWLEDGEMENTS

I love science and working in the lab. I would however never have developed such an attitude towards my work if there were not so many people supporting me. First, I would like to express my gratitude to Prof. Dr. Thierry Hennet, for providing the possibility to work in his lab, for great supervision and for fruitful discussions not only about scientific issues but also on whisk(e)y, photography and many things more. I thank Thierry for his patience and the continuous support and mentoring during my whole PhD.

I further want to thank my PhD committee, namely Prof. Dr. Lubor Borsig, Prof. Dr. Martin Hersberger and Prof. Dr. Richard Steet for guidance through my thesis and helpful advices during committee meetings.

During the last years, I worked together with many people who turned into friends. It was a pleasure to work with all of you. I would especially like to say thanks to Michi, Jürg, Anna, Adrienne, Nikunj, Eddie, Jesus, Marek, Nina and Darya for scientific discussions and advices, for glorious times at the hat-making sessions, at going out and after-work beers and especially during our various trips to Dublin, Amsterdam and to the States resulting in memories I will never forget. I would also like to thank current and former members of the Hennet and Borsig groups, namely Kelvin, Andreas, Katharina, Sacha, Luca, Tom, Christoph, Irina, Katja, Alexa, Marko, Cristina and Yee Ling for establishing a productive lab environment and provision of a helping hand if needed. The same is true for the whole L-Floor, especially the Devuyst Group. Thanks a lot for organization of the SoLa, for summer BBQ's, Halloween parties and much more. Special thanks belong also to Giovi for her care and assistance with all issues related and unrelated to the studies, for her awesome Lasagne and the organization of the hat-making sessions.

I want to express my profound gratitude to my parents, Simone and Jessi for their great support during all times of my studies, for believing in me and for their kind words of motivation when needed.

6. CURRICULUM VITAE

Education and Qualifications

05/2011 – 03/2016	**University of Zurich, Institute of Physiology,** *Ph.D. Studies in integrative molecular medicine in the lab of Prof. Dr. Thierry Hennet*
09/2009 – 10/2010	**Swiss Federal Institute of Technology (ETH)**, *Master of Science ETH in Micro- and Immunbiology in the lab of Dr. Nicola Zamboni*
09/2005 – 09/2009	**Swiss Federal Institute of Technology (ETH)**, *Bachelor of Science ETH in Biology with focus on Chemistry*
08/1999 – 09/2005	**High school Zurich Oerlikon, A level** *Focus on Ancient Greek and Latin (Type A)*

Publications and Conference Participations

2016:	**Collagen accumulation in osteosarcoma cells lacking GLT25D1 collagen galactosyltransferase** *Stephan Baumann, Thierry Hennet, Submitted to Journal of Cell Science*
2016:	**Collagens in giant viruses** *Stephan Baumann, Nina Hochhold, Thierry Hennet, in preparation*
2013:	**Recombinant expression of hydroxylated human collagen in Escherichia coli** *Christoph Rutschmann and Stephan Baumann, Jürg Cabalzar, Kelvin B. Luther, Thierry Hennet in Journal of applied microbiology and biotechnology*
2011:	**Engineering genetically encoded nanosensors for real-time in vivo measurments of citrate concentrations** *Jennifer Christina Ewald, Sabrina Reich, Stephan Baumann, Wolf B. Frommer and Nicola Zamboni in PLoS ONE*
2013:	***Gordons Research Conference in Glycobiology*** *High level expression of posttranslationally modified collagen in E. coli, Ventura, CA, USA*